TITLE:

Studies on the analysis and recognition of Japanese speech sounds(Dissertation_全文)

AUTHOR(S):

Doshita, Shuji

CITATION:

Doshita, Shuji. Studies on the analysis and recognition of Japanese speech sounds. 京都大学, 1966, 工学博士

ISSUE DATE:

1966-09-27

URL:

https://doi.org/10.14989/doctor.k622

RIGHT:

京都大学学術情報リポジトリ
KURENAI 紅
Kyoto University Research Information Repository

新制
工.
80

STUDIES ON THE ANALYSIS AND RECOGNITION OF JAPANESE SPEECH SOUNDS

By

SHUJI DOSHITA

Thesis, Kyoto University

September, 1965

STUDIES ON THE ANALYSIS AND RECOGNITION OF JAPANESE SPEECH SOUNDS

By

SHUJI DOSHITA

Thesis, Kyoto University

September, 1965

PREFACE

Speech sound is as important a means of human communication as letter. The research on speech has been tried for many years in several fields such as engineering, medical science, linguistics and psychology. In recent years much more development has been made on the mechanical recognition and generation of speech, based on the development of electronics and information processing technique. The mechanical processing of speech sound is not only of interest to the engineering field, but it also contributes to other fields. On the other hand the knowledge of the other fields, such as hearing mechanism and speaking mechanism of human beings, structure of phonetics and linguistics and human mechanism of perception and recognition, contributes much to this field.

The recognition of speech has a close relation with pattern recognition problem based on general principle, having the self-organizing or learning mechanism. The speech sound, however, is so complicated that we can not clarify the mechanism of speech by the general principle without giving any concrete structure or rules in its recognition system. In the course of recognition of speech, it is a problem of primary importance how to give the sufficient inputs or parameters for the recognizing system and how to give basic rules to it. Thus, with the approach by general principle, the analysis of speech sound properties and the study on the structure of the phonetics and the linguistics are more and more important.

Such development of the speech recognition and production, together with character recognition and machine translation, will make it possible to construct linguistic automata which automatically perform the processing of the linguistic informations that are now done by human beings. Also, it will make it possible for the present computers to have the function of automatic transfer of information with the external world.

The speech sound is characterized by the fact that it is articulated under the constraints of the speech organs. The phonetic structure of the speech sound is formed in relation with the characteristics of hearing organ and the phonemic system of the particular language. To master the speech system human beings make feedback loop by hearing organ and articulatory organ, referring to the speech sounds of the other speaker. In the utterance, the articulatory organ operates under the control of the loop.

In the same way it may be thought that, in mechanical processing system of speech, the recognition of speech must be solved first and, by using these results, the synthesis system may be formed by the analysis-synthesis loop. Recognition of speech sound may have two aspects: the analysis and the determination or discrimination. The analysis is to form the parameter space of the input speech, corresponding to the perception of tonal quality of speech sound in human beings. The processing of speech in this level is essentially the problem concerning the acoustic properties of speech sound wave and the analyzing network of hearing organ. It is, therefore, necessary to find such sufficient parameters, comparable with the ability of human beings, that can describe the speech sound and the speech elements constituting the speech sound, i.e., the monosyllable or phoneme.

In the connected speech which is the sequence of such elements, the sound wave is not the simple connection of each elementary sound wave, but, by the co-articulation effect between phonemes, they are affected each other. The primary effect is, however, considered to occur between adjacent phonemes. Therefore, by taking the three phoneme sequences as the basic recognition element, the connected speech can be treated under the general principle without imposing the restriction of operation. On the other hand, the segmentation is necessary to separate the sound wave into elementary sections corresponding to the phoneme.

The above processing is principally the problem of speech sound

itself, i.e., of the acoustic level and the phonetic level. Thus we can recognize the series of phonemes from the speech sound without considering its contents of information. This is the conversion from the "speech sound to letter or code sequence", where the input is regarded as the random sequence of phonemes, generated under the rules of phonetics. As concerns the sound, this conversion may be sufficient.

In the conversational speech sound, the final goal of recognition is the "conversion from speech sound to sentence" or the understanding of the message. This can be done by utilizing the linguistic structure, syntax and semantics, on which many researches have been performed in relation to machine translation. The daily conversational speech sound is not always articulated so clearly as one pronounces the random sequence of phonemes, but, utilizing much redunduncy of linguistic structure, articulation is insufficient for the perfect recognition of letters by the processing up to the phonetic level. The uncertainty may be solved by the processing at the linguistic level.

From this point of view the primary problem of speech recognition may be considered as the conversion from speech sound to the sequence of phonemes, where the machine is required to have the ability to recognize perfectly the connected speech sound which has no linguistic structure (syntactic and semantic structure) and which human auditory system can understand.

In this paper the conversion from speech sound to phoneme sequence is described, that is, the acoustic processing and the phonetic processing. The description is divided into two parts. In PART I analysis of speech sound by filter bank and the statistics of the conversational speech sound are described. And in PART II automatic recognition system of the Japanese connected speech sound, which is the special purpose machine designed to recognize it in real time, is discussed, together with zero-crossing analysis.

(As for PART I:)

In chapter 2 the response of the single tuned filter to formant signal was examined, using the envelope detection by condenser input type smoothing network, to give the estimation of its response to speech sound.

In chapter 3 the description of the spectrum analyzer using the single tuned filter bank and that of the display device are given. They were designed to be able to obtain detailed representation on the spectrum of speech sound. The responses of those devices to several signals such as sine wave, white noise, impulse train, formant-shaped signal, etc. are also discussed.

The analysis of Japanese monosyllables and connected speech sound is described in chapter 4. The analysis was performed using the analyzer stated in chapter 3. The results are summarized in APPENDIX I of the supplements (in annexed volume).

In chapter 5 the statistical properties of Japanese phonemes are investigated. The results were summarized as trigram, digram, entropy, etc., which gave the knowledge for the design of conversational speech recognition system.

(As for PART II:)

In chapter 2 the phase characteristics of the signals are treated. After the discussion on the representation of the zero-crossing signal, the analysis circuits to obtain the zero-crossing distribution are presented, which are used as the analysis circuit of speech recognition system in chapter 4 and 5. The zero-crossing analysis method combined with single tuned filter is proposed. It was applied to the formant extraction of speech sounds.

In chapter 3 an automatic recognition system of Japanese speech sounds is discussed, which performs the segmentation by detecting the "stability" and recognizes the connected speech sounds.

In chapter 4 the speech pattern recognition system is discussed that can process the speech sound in real time considering the phonetic contextual effects. A new representation of the speech pattern by a combination of sequence pattern and weight pattern is proposed. The principle was applied to the vowel pattern recognizer which was combined with the speech recognition system stated in chapter 3.

CONTENTS

PART II

AUTOMATIC RECOGNITION OF JAPANESE SPEECH SOUNDS

Chapter

SUPPLEMENTS (ANNEXED VOLUME)

PART I

ANALYSIS OF JAPANESE SPEECH SOUNDS

Chapter 1

INTRODUCTION

Speech sound is the sound wave generated by the human articulating organ. By this fact a strong restriction is given in the structure of speech sound which distinguishes it from natural sound. Speakers want to send the information to the listener through it. One of the principal informations is the systematic linguistic information which represents the symbolic information. Prosody and naturalness are the other kinds of informations particular to the speech sound which are not translated into letters in one to one correspondence. The acoustic information of the speech sound wave is associated with the linguistic representation by the phonetic structure which is particular to each language. In most cases, recognition of speech means the recognition of the linguistic information.

Corresponding to the mechanism of the speech production, the processing of speech is considered in several levels. First the speech sound wave must be analyzed to obtain the physical parameters. Next, according to the phonetic structure, the parameters are associated with phonemes or elementary units of the language. Through these steps we can find the sequence of phonems or letters from the input speech sound. The final level is the

linguistic processing to know the meanings of the message that the speech sound conveys. In this paper the acoustic properties and the phonetic aspects of the speech sound are examined.

The speech sound wave is the sound pressure outputs from transmission networks excited by several types of sources as shown in Fig. 1.1.[1] The transmission networks are the pharynx cavity, the mouth cavity and the nasal cavity, which are approximated as one dimensional acoustic tubes

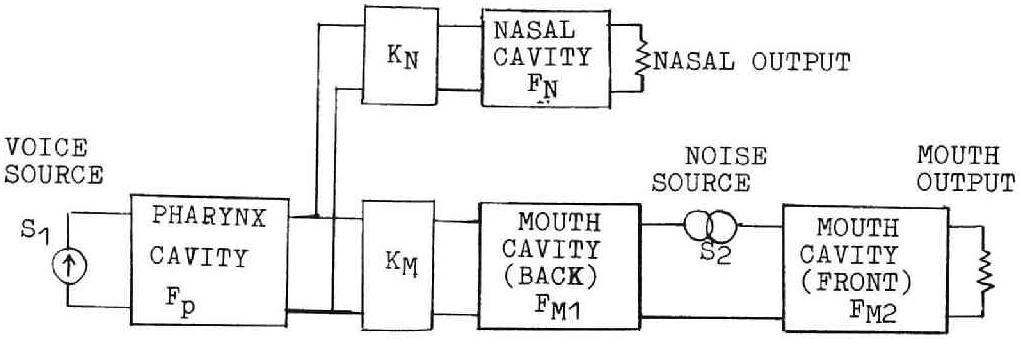


Fig. 1.1 Block giagram of speech production.

having the cross-sectional area depending on the position along the axis. Each part of the mouth cavity area changes slowly (or abruptly in some time points), resulting a time varying transfer function. The shape of the vocal tract is characterized by the place of articulation at which the constriction of cross-sectional area occurs.

The output of the pharynx cavity is connected to the mouth cavity and the nasal cavity by the coupling factor K_N and K_M, where K_M+K_N is nearly constant. For $K_M=0$, nasal sound is generated and for $K_N=0$, non-nasal sound. In nasalized vowel K_N and K_M have appropriate values other than zero, resulting the parallel outputs from the mouth cavity and the nasal cavity.

There are several types of sources situated at several positions of

the vocal tract. Vocal cords S_1, situated behind the transmission cavity makes relaxation oscillations of volume velocity and generates pulsive wave forms with roughly constant period. Another type of source is the noise source of constant pressure type by the constriction at some position of the mouth cavity. By the narrow constriction turbulent noise is generated and, by the abrupt opening of some section of the tract, transient random impulses are generated. In these cases the back cavities behind the source can act as the anti-resonance networks. In burst type source, the transfer function of the vocal tract also makes abrupt change, generating a complicated sound pressure.

In vowel, the source is the voiced source and the nasal cavity is not coupled. The transfer function is attained by the cascade connection of cavities F_P, F_{M1} and F_{M2}, resulting some resonance characteristics (formants). The tone quality of a vowel sound is characterized by the shape of the mouth cavity.

In unvoiced consonants the source is the noise scurce situated in the mid point between F_{M1} and F_{M2} where the mouth is narrowed. Nasal cavity is not coupled. By this source position, along with the formants, some anti-formants (anti-resonances) by back cavities F_P and F_{M1} exist in the output. When S_2 is turbulent noise it is a stationary noise sound and, when S_2 is transient random noise it is a stop consonant. For the voiced consonant two sources S_1 and S_2 operate at the same time, the operation of S_2 being modulated by the vocal cord vibration.

Among those factors that characterize speech, the factor corresponding to the transfer function of the vocal tract and the source position of noise source are called the "place of articulation" and the factors corresponding to the source type and to the nasal coupling are called the "manner of articulation". According to the latter factors speech sounds are classified into several groups such as; vowel-like sound, unvoiced consonant,

voiced consonant, nasal consonant, stop consonant, noise consonant, etc., each of which has the distinctive differences in physical properties of the speech sound. To carry out analysis these differences must be considered.

The analysis of speech sound is to find the physical parameters that describe the phonetic structures. Many kinds of analyses have been tried such as time domain analysis, frequency analysis, orthogonal function expansion, etc.. From the analytical point of view the speech sound is non-stationary and is the succession of signal segments having different properties, each of which represents the linguistic informations to be transmitted. Therefore, the method of analysis must reserve time variation of speech parameters. In that sense the parameters are called "short-time" or "running".

The most popular method of analysis is the frequency analysis by band pass filters. The result of this method given as a set of filter responses makes a short-time power spectrum presented as time, frequency and intensity. Any methods may not be general, because the results are subject to the adopted method such as: filter attenuation characteristics, band widths of filters, arrangement of center frequencies of the band pass filters and further the method detecting the amplitude of responses. There are two types to realize the device; one is the heterodyne type like Sonagraph which sweeps a filter step by step within the frequency range of interest and the other is the real time analyzer of bank-of-filter type which prepares a bank of filters covering the range. The device used in this paper is the semi-heterodyne analyzer, such as Sonalator, by a bank of filters which analyzes the overall range in several times by shifting the signal frequency.

In earlier days the channel number of filters was not so many[2][3] and the sharp cut off filters were used. Recently the filtering problems in the analysis have been discussed. The discussions were mainly made on

the frequency responses of the several types of filters to the periodic signal (vowel-like sound) and the stationary noise (consonant).[4][5] For voiced sound the problem is the fluttering of spectrum envelope by the mutual action among the fundamental frequency, the filter band width and the filter center frequency and is the response to the formant. In general there are two types of approaches for the analysis of the voiced sound: to obtain the smoothed spectrum envelope and to obtain the amplitude of each harmonic component separately. In the device of chapter 3 the harmonic structure is analyzed by rather narrower band width filters. As well as the frequency response, the time response is important which is discussed in chapter 2.

One of the measures of a filter is the frequency resolution $\Delta\omega$ and time resolution Δt defined as follows:[6]

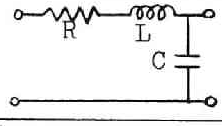
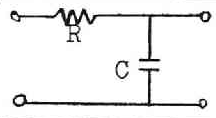
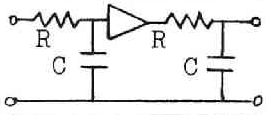
$$(\Delta t)^2 = \frac{\int_{-\infty}^{\infty} t^2 |f(t)|^2 \, dt}{\int_{-\infty}^{\infty} |f(t)|^2 \, dt} \qquad f(t) : \text{Impulse response}$$

$$(\Delta\omega)^2 = \frac{\int_{-\infty}^{\infty} (2\pi f)^2 |g(f)|^2 df}{\int_{-\infty}^{\infty} |g(f)|^2 \, df} \qquad g(f) : \text{Spectrum}$$

A merit of filter is given by $K = \Delta t \cdot \Delta\omega \geqslant 1/2$, in which equality is valid for Gaussian type filter.[6] Table 1.1 shows these values for some types of filters.

The sharp cutoff filter is good for the frequency resolution and the broad characteristic filter is for the time response. From these points Gaussian filter is considered to be suitable to speech analysis.[7][8] For the single tuned filter, K becomes infinite by the lacking of frequency resolution. However, the response to the speech especially for vowel sound must be discussed from the other point, because the speech sound has the formant structure. The fact that the transient time is small and the rise time is zero means the quick response to the transient impulse such as

Table 1.1 Transfer functions $F(j\omega)$, time resolution Δt, frequency resolution $\Delta\omega$ and figure of merit K of several types of filters.

	TYPE OF FILTER	$F(j\omega)$	Δt	$\Delta\omega$	$K=\Delta t\cdot\Delta\omega$
1		$\frac{1}{1+j2\zeta^{*}x-x^{2}}$ $x=\omega/\omega_c$**	$\frac{\sqrt{4\zeta^4-\zeta^2+1}}{2\omega_c}$	ω_c	$\frac{\sqrt{4\zeta^4-\zeta^2+1}}{2}$
2		$\frac{1}{1+j\omega Q}$ $Q=CR$	$\frac{Q}{\sqrt{2}}$	∞	∞
3		$\frac{1}{(1+j\omega Q)^2}$ $Q=CR$	$\sqrt{3}Q$	$\frac{1}{Q}$	$\sqrt{3}$
4	RECTANGULAR TYPE	$1\ (\lvert\omega\rvert\leq\omega_p)$ $0\ (\lvert\omega\rvert<\omega_p)$	∞	$\frac{\omega_p}{\sqrt{3}}$	∞
5	GAUSS TYPE	$e^{-k\omega^2}$	$\sqrt{k}$	$\frac{1}{2\sqrt{k}}$	$\frac{1}{2}$

* $\zeta=\frac{R}{2L\omega_c}$ ** $\omega_c=\frac{1}{\sqrt{LC}}$

caused by the burst of consonant. For this reason the single tuned filter was adopted.

Another problem in filter analysis is the smoothing network to detect the envelope of response. For the analysis by the sharper cutoff analyzing filter in which the emphasis is put on the frequency response, the low pass filter of sharp cutoff characteristics was used to smooth out the fluctuation by the harmonic structure. In such case the instantaneous time response of analyzing filter is smoothed out, too. As, for the broad characteristic analyzing filter in which time response must be regarded, the response forms the damped oscillation, the condenser input type CR smoothing circuit may be preferable.

Along with the filter analysis, there are several types of frequency analysis such as short-time-spectrum analysis by auto-correlation method,

[9][10] real time short-time-spectrum analysis by recirculating delay line, [11] etc..

To process the analyzed results by computers and to display them as patterns or spectra, a set of filter outputs must be sampled and multiplexed to a serial signal. To retain the transient response, the sampling speed must be faster than required by the sampling theory.

The speech sound is represented as the cascade of transfer functions of such as source, formants and anti-formants, etc., though there are some exceptions in the nasalized vowel sound in which the output is the sum of outputs of the mouth cavity and the nasal cavity. Therefore the amplitude representation[12][13] in logarithmic scale is used here.

Extensive studies on the analysis by band pass filters have been worked using the Sonagraph.[14] But it takes a fairly long time and is not suited for the direct application to parameter extraction system. The device presented here has the comparable ability with Sonagraph and can get results in a shorter processing time, without disregarding the essential time response which is neglected in the ordinary spectrum analysis.

The basic unit of articulation is monosyllable in which almost essential parameters are contained. Some parameters correspond to the manner of articulation, the others to the place of articulation. In the following chapter the qualitative features on the distinction between phoneme groups such as voiced/unvoiced, nasal/non-nasal, etc. are studied. Thousands of spectrum patterns or sections were recorded on the photographic film, some of which are shown in APPENDIX I of the supplements.

In the connected speech which is composed of a sequence of phonemes, the sound is not a simple connection of monosyllabic sounds, but there are some intereffects between phonemes by co-articulation effects in phonetic level. In the latter part of chapter 4, the analysis of connected phoneme sequences was performed. The materials (words) which contain the

phoneme sequence to be examined were selected systematically, so that the differences between them are contrasted each other.

On the other hand, in the recognition of conversational speech, it is necessary to examine the linguistic structure of the language of interest. In the processing of phonetic level the statistical structure on the phonetic sequence is needed. By selecting three phoneme sequences as the elementary recognition unit, the recognition system of conversational speech sound was organized (refer to PART II). In chapter 5 the statistical properties of the Japanese phoneme sequence are examined.

Chapter 2

RESPONSE OF SINGLE TUNED FILTER TO FORMANT

2.1 Selection of Type of Analyzing Filter

The results of analysis by bank of filters depend on the characteristics of analyzing filters and also on whether the emphasis is put on the frequency response or the time response. The selection of filter characteristics is made from various stand points such as frequency response[5], time response[15], simulation of mechanism of the ear.[16] When the frequency response is considered as important, the band pass filter of sharp cutoff characteristics is used, followed by the smoothing low pass filter of lower cutoff frequency after the rectification, yielding rather averaged envelope. On the other hand the time response must be considered in analysis, since the transient characteristic is also one aspect of speech sound. The response of narrow band, sharp cutoff band pass filter is almost decided by the filter itself and the transient characteristics of speech sound are not always reflected in the response. To let the transient characteristics reflect on the response, the filter having the characteristics near to those of the speech sound may be desirable. Refering to the source characteristics, the speech sound may roughly consist of the signal with quasi-periodic property by the vocal cord vibration, characterized by formants, the stationary or modulated noise signal and the transient noise (burst). The time response of filter to each type of the signals may be an interesting problem. As the most parts of the speech sound have the formant structure, the broad cutoff filter may be recommended as analyzing filter. Here the single tuned filter was adopted.

2.2 Response to Formant[19]

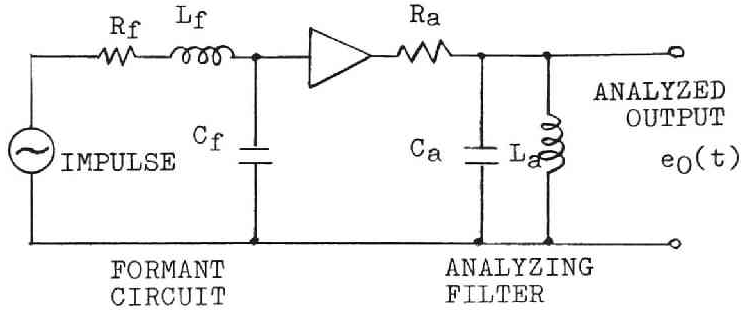


Fig. 2.1 Circuit to calculate the response of single tuned filter to formant.

The disadvantage of single tuned filter is the lacking of the frequency resolution by its gradual cutoff characteristics[4], though it has the faster rise time and simple time response. One of the purposes of the analysis by single tuned filters is to examine its time response as well as the frequency response to the speech signal having different time characteristics. In this chapter the responses to the formants are examined.

The analyzing filter is realized by a parallel resonance circuit as shown in Fig. 2.1. The transfer function is

$$G_a(s) = \frac{1}{C_a R_a} \frac{s}{(s+\sigma_a)^2 + \omega_a^2} , \qquad (2.1)$$

in which

$$\sigma_a = \frac{1}{2R_a C_a} ,$$

$$\omega_a = \sqrt{\frac{1}{L_a C_a} - \frac{1}{4R_a^2 C_a^2}} .$$

The impulse response is expressed as

$$g_a(t) = Ae^{-\sigma_a t} \cdot \sin(\omega_a t + \varphi_a) . \tag{2.2}$$

The speech signal has several formants. The response to formant is discussed below.

The formant is realized by a serial resonance circuit as in Fig.2.1, having the transfer function

$$G_f(s) = \frac{1}{L_f C_f} \cdot \frac{1}{(s+\sigma_f)^2 + \omega_f^2} . \tag{2.3}$$

The output is connected to the analyzing filter. The output $E_o(s)$ when driven by a single impulse is

$$E_o(s) = K \frac{s}{\left\{(s+\sigma_f)^2 + \omega_f^2\right\}\left\{(s+\sigma_a)^2 + \omega_a^2\right\}} , \tag{2.4}$$

in which: $K = \dfrac{1}{L_f C_f C_a R_a}$,

$$\sigma_f = \frac{R_f}{2L_f} = \pi B_f , \qquad \sigma_a = \frac{1}{2C_a R_a} = \pi B_a ,$$

$$2\pi F_f = \omega_f = \sqrt{\frac{1}{L_f C_f} - \frac{R_f^2}{4L_f^2}} , \qquad 2\pi F_a = \omega_a = \sqrt{\frac{1}{L_a C_a} - \frac{1}{4C_a^2 R_a^2}} .$$

The time function of the response is;

(i) when $\sigma_f + j\omega_f \neq \sigma_a + j\omega_a$,

$$e_o(t) = \mathcal{L}^{-1}\left[E_o(s)\right] = K' \sum_i \sqrt{1+\delta_i^2}\, e^{-\sigma_i t} \sin(\omega_i t + \varphi_i) ,$$

$$\text{for } i = f, a \tag{2.5a}$$

(ii) when $\sigma_f + j\omega_f = \sigma_a + j\omega_a = \sigma + j\omega$,

$$e_o(t) = \frac{K}{2\omega^2} e^{-\sigma t} \cdot \left(\omega t - \frac{1}{\alpha}\right)\left\{ \sin \omega t - \frac{\omega t}{1-\alpha\omega t} \cos \omega t \right\} , \tag{2.5b}$$

in which:

$$K' = \frac{K}{\sqrt{a_f^2 + b_f^2}} = \frac{K}{\sqrt{a_a^2 + b_a^2}} ,$$

$$a_i = (\sigma_a - \sigma_f)^2 + \omega_j^2 - \omega_i^2 ,$$

$$b_i = 2\omega_i (\sigma_j - \sigma_i) ,$$

$$\varphi_i = \tan^{-1} \frac{a_i\omega_i + b_i\sigma_i}{-a_i\sigma_i + b_i\omega_i} , \qquad (i, j = a, f \ ; \ i \neq j)$$

$$\zeta_i = \sigma_i/\omega_i , \qquad \alpha = \omega/\sigma .$$

The response $e_o(t)$ may depend on the frequency F_f and the band width B_f of the formant and the center frequency F_a and the band width B_a of the analyzing filter. As in most cases B_f and R_a are considerably smaller than F_f and F_a, respectively, (2.5a) and (2.5b) can be represented by the amplitude and the phase component.

$$e_0(t) = a_0(t) \sin \phi_0(t).$$

For (2.5a), the amplitude is:

$$a_0(t) = K' \sqrt{(1+\zeta_f^2)e^{-2\sigma_f t} + (1+\zeta_a^2)e^{-2\sigma_a t} + 2\sqrt{(1+\zeta_f^2)(1+\zeta_a^2)}} \times e^{-(\sigma_f + \sigma_a)t} \times \cos\{(\omega_a - \omega_f)t + \varphi_a - \varphi_f\}. \qquad (2.6a)$$

(The phase characteristics will be discussed in chapter 2, PART II in relation to the zero-crossing analysis of filtered signal.)

For the condition $\zeta_f \ll 1$ and $\zeta_a \ll 1$,

$$a_0(t) \doteqdot K' e^{-\sigma_f t} \sqrt{1 + e^{-2\Delta\sigma t} + 2e^{-\Delta\sigma t} \cos(\Delta\omega t - \pi)} \qquad (2.6a')$$

in which $\Delta\omega = \omega_a - \omega_f$, $\Delta\sigma = \sigma_a - \sigma_f$.

For (2.5b) the amplitude is

$$a_0'(t) = \frac{K}{2\omega^2} \cdot e^{-\sigma t} \sqrt{(\sigma^2 + \omega^2)t^2 - 2\sigma t + (\sigma/\omega)^2} . \qquad (2.6b)$$

From (2.6a) and (2.6b) it is expected that the ratio of $a_0'(t)$ to $a_0(t)$ becomes large with time. In (2.6a), $a_0(t)$ oscillates at the frequency $\Delta\omega/2\pi$. The envelope $b_0(t)$ of the amplitude $a_0(t)$ is given from the condition $\cos(\Delta\omega t+\varphi_a-\varphi_f)=1$:

$$b_0(t) = K'e^{-\sigma_f t}\left(\sqrt{(1+\zeta_f^2)}+\sqrt{(1+\zeta_a^2)}e^{-\Delta\sigma t}\right),$$

for $\zeta_f \ll 1$, $\zeta_a \ll 1$

$$b_0(t) \doteqdot K' e^{-\sigma_f t}\left(1+e^{-\Delta\sigma t}\right), \qquad (2.7)$$

and for $\sigma_a = \sigma_f = \sigma$, $\omega_a \neq \omega_f$

$$b_0(t) \doteqdot K'e^{-\sigma t} .$$

At $t=0$,

$$b_0(0) = K' = \frac{K}{\sqrt{(\Delta\sigma)^4+(\omega_a^2-\omega_f^2)^2+2(\Delta\sigma)^2(\omega_a^2+\omega_f^2)}} , \qquad (2.8\ a)$$

$$\Delta\sigma = \sigma_a - \sigma_f$$

for $\sigma_a=\sigma_f=\sigma$, $\omega_a \neq \omega_b$

$$b_0(0) = \frac{K}{\left|\omega_a^2-\omega_f^2\right|} . \qquad (2.8\ b)$$

From these relations, it is seen that the envelope of amplitude significantly changes with σ_a, σ_f, ω_a and ω_f and its value is large when the center frequency of analyzing filter comes close to the formant frequency.

To examine the response of filter bank composed of single tuned filters to formant, amplitude response (2.6a) and (2.6b) were calculated. Fig.2.2 and Fig.2.3 show the amplitude response to the formant of F_f=1000cps corresponding to $\sigma_a=\sigma_f$ and $\sigma_a\neq\sigma_f$, respectively. The abscissa is time after the impulse was applied and the ordinate is the relative amplitude. In Fig.2.2, a) is the case of $B_a=B_f$=100cps, b) is $B_a=B_f$=33cps. In case of $B_a=B_f$, the $b_0(t)$ is straight line as shown

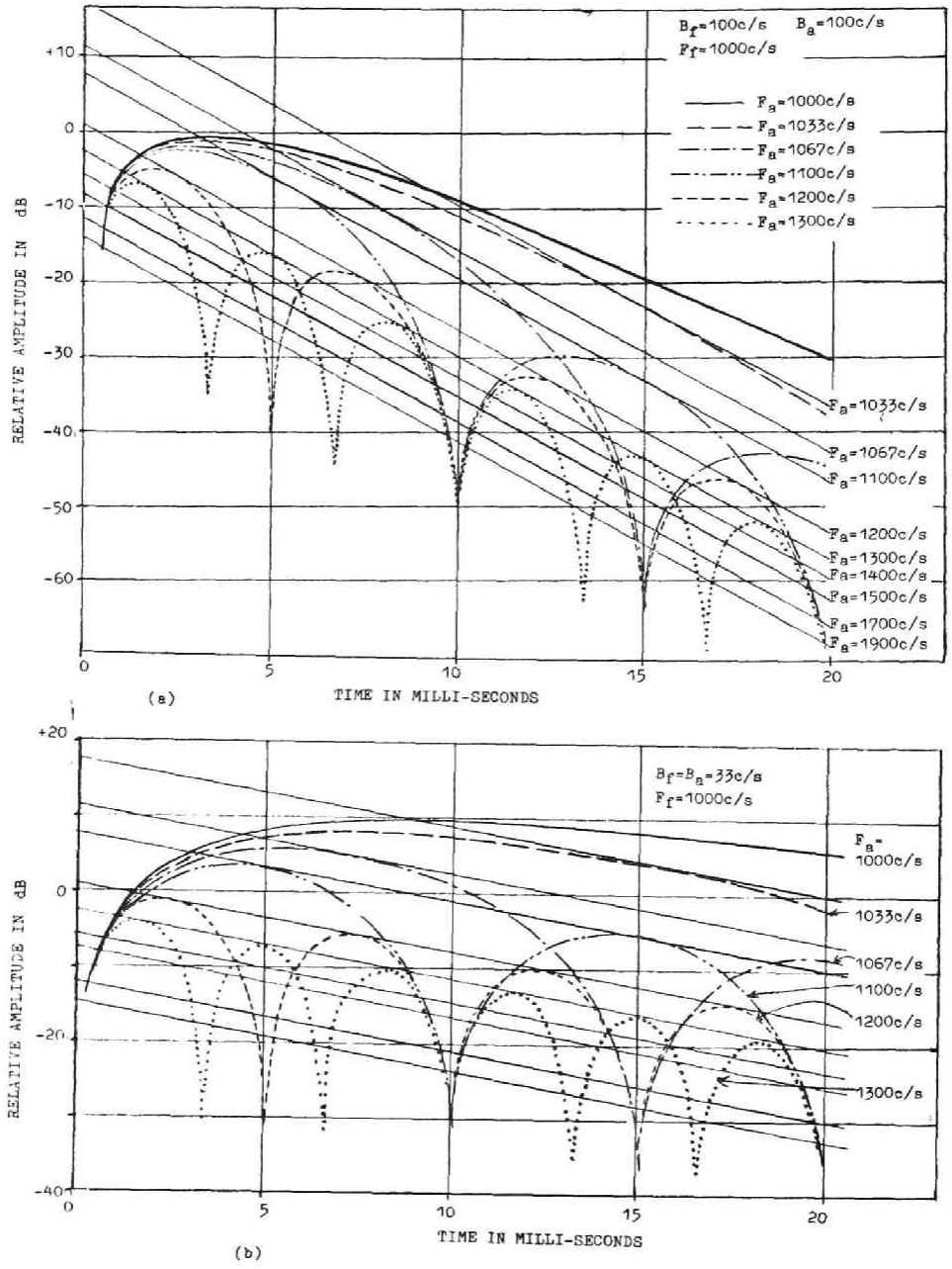


Fig. 2.2 Amplitude response of single tuned filters with different center frequencies to formant. The straight lines show $b_0(t)$'s of equation (2.7). Abscissa is the time after the impulse is applied, ordinate is the amplitude response. The parameter is center frequency of analyzing filter.

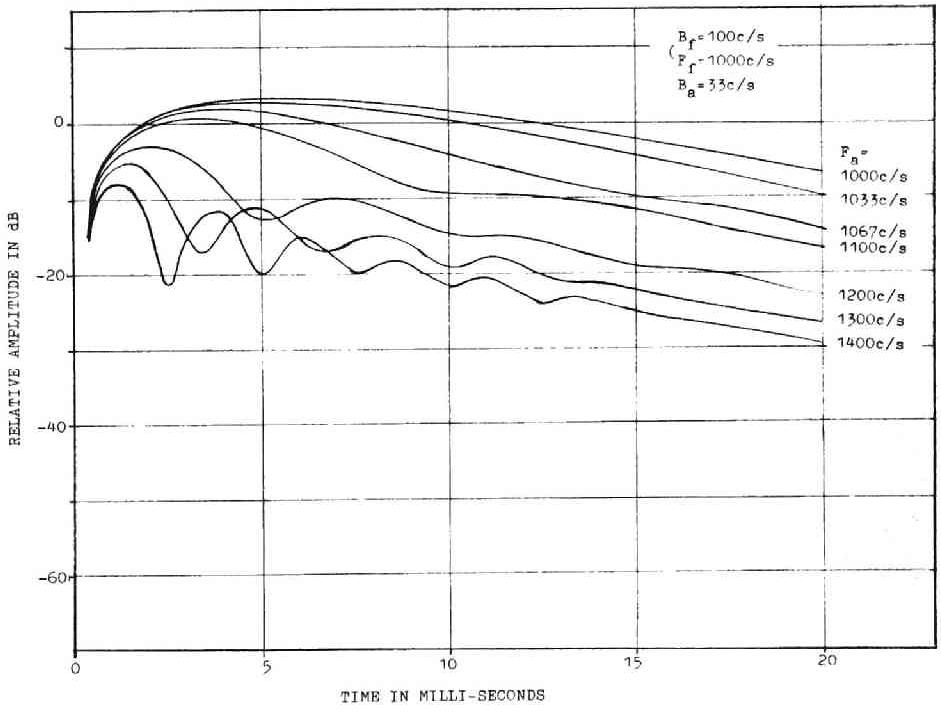


Fig. 2.3 Amplitude response of single tuned filters with different center frequencies to formant for the condition that band widths of filter and formant are different. Abscissa and ordinate are the same as Fig. 2.2.

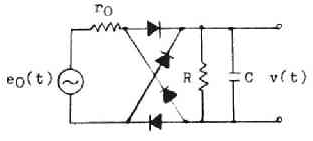


Fig. 2.4 Rectification circuit with RC network for the smoothing of the single tuned filter response.

in the figure whose gradient is given by σ and which is expressed as

$$b_o(t) = \frac{K}{|\omega_a^2 - \omega_f^2|} e^{-\sigma t} \qquad (\omega_a \neq \omega_f). \tag{2.9}$$

The actual amplitude will oscillate in such way to contact with $b_o(t)$ having the sharp depression at energy $2\pi/\Delta\omega$ sec. In the case of $B_a \neq B_f$ (Fig.2.3), the oscillation is not so sharp and its amplitude decreases at the order of $\Delta\sigma$.

2.3 Detection of Envelope of Amplitude Response

As stated above the amplitude response of a single tuned filter to a formant shows complicated apperance which will suggest the necessity to select the appropriate type of smoothing circuit in the realization of frequency analyzer.

One of the possible methods is to use a low pass filter after the rectification so that it can smooth the fluctuations. In this method, however, the transient response is smoothed, too. The alternative method is to use the rectifier followed by CR network or the condenser input type rectifier, with appropriate time constant. In Fig.2.4 the response of analyzing filter $e_o(t)$ is rectified and the envelope v(t) is detected. The source resistance r_o is selected as

$$r_o \ll R ,$$

yielding fast rise and slow decay.

The charging time constant $\tau_f = Cr_o$ is chosen so small that the envelope v(t) can follow the building up of $a_o(t)$ or $a'_o(t)$ expressed by equation (2.6a) or (2.6b). The discharging time constant $\tau_- = CR$ is chosen to be able to smooth the fluctuation of $a_o(t)$. The fluctuation is very remarkable in case of $\sigma_a = \sigma_f$. Therefore, by using the value τ_- equal to the time constant σ_a of the analyzing filter, the approximate envelope of $a_o(t)$ may be

obtained. The condition is not completely satisfied for all the speech signals, in which case, however, the fluctuation is not so large. Thus, the output v(t) rises following to the rise of $a_o(t)$ and after $a_o(t)$ reached to $b_o(t)$, it approximately follows to $b_o(t)$. Fig.2.5(a) and (b) show the response v(t) for $B_f=B_a=100$ cps and for $B_f=B_a=33$cps, respectively, in which τ is ideally chosen.

2.4 Response of Filter Bank

The amplitude response or the output v(t) stated above will have different tendency according to the parameters such as B_f, B_a, F_f and F_a. Since one of the purposes of this experiment is to detect the formant frequency, the response is desired to have a large value for $F_a=F_f$. For the responses of Fig.2.5, i.e., for the condition of $B_f=B_a$, these relations are stated below.

For $\omega_a \neq \omega_f$, at the time

$$t_1 = \frac{\pi}{|\omega_f - \omega_a|} \quad \frac{1}{2|F_f - F_a|}$$

the amplitude response reaches $b_o(t)$. The response of filter, whose center frequency is matched with that of the formant, i.e., $F_a=F_f$, is approximated for $t \gg t_2=\frac{2\sigma}{\omega^2}$ as

$$a_o'(t) \doteqdot \frac{K}{2\omega_f} e^{-\sigma t} \cdot t \qquad (2.10)$$

The ratio of the envelope responses of filters having the center frequency $F_a \neq F_f$ to that of filter having the center frequency $F_a=F_f$ is given by

$$R = 20 \cdot \log \frac{a_o'(t)}{b_o(t)} = 20 \log t - 20 \log \frac{|\omega_a^2 - \omega_f^2|}{2\omega_f} \qquad (2.11)$$

for the time that satisfies $t>t_1$ and $t>t_2$.

To the filter $F_a \doteqdot F_f$, (2.11) is approximated as

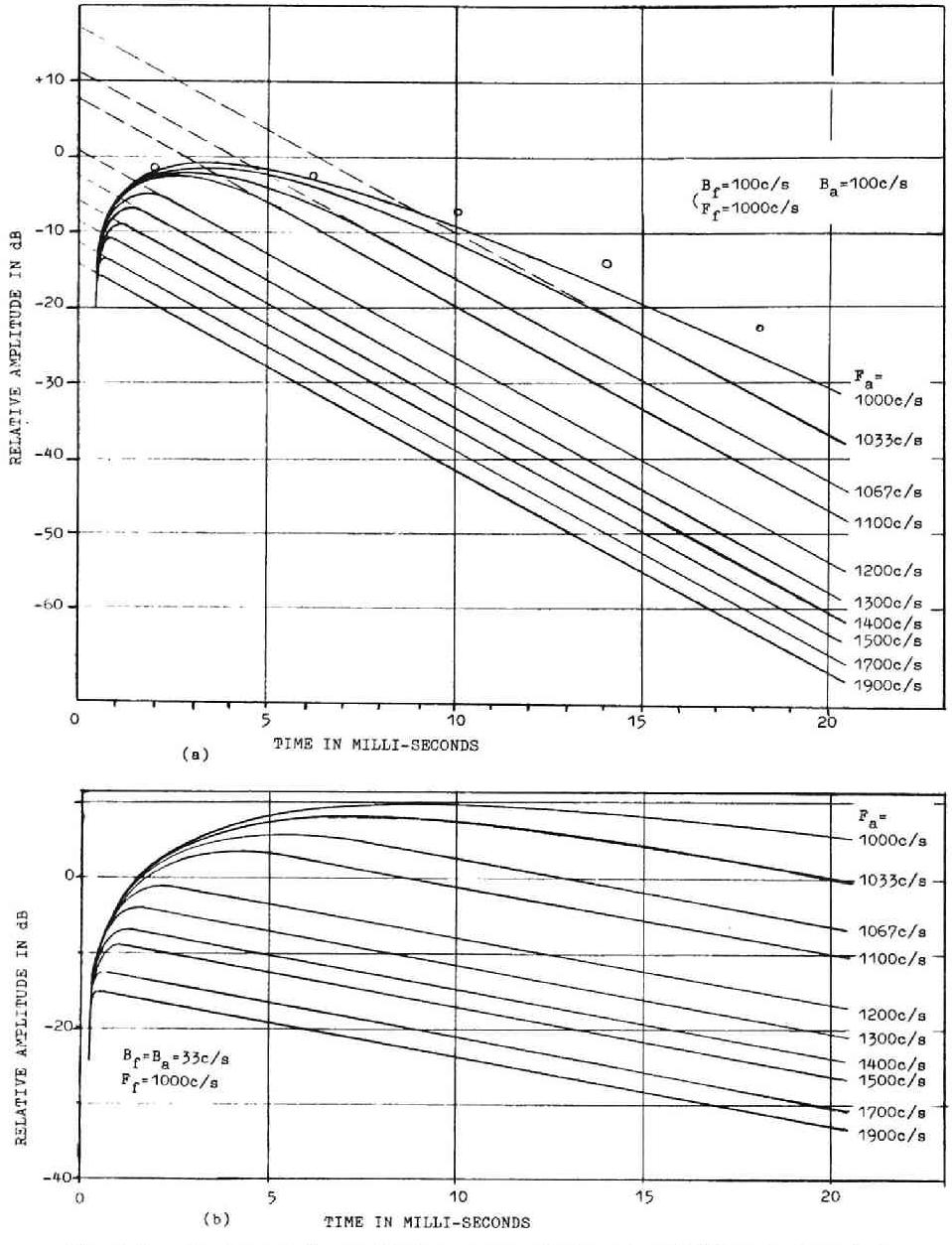


Fig. 2.5 Envelope of the amplitude response of single tuned filters to formant for the condition B_f=B_a, rectified by the circuit of Fig 2.4 with the ideally chosen discharging time constant CR. Abscissa is time after the impulse. Open circles in Fig. (a) show the data obtained from analyzing filter of chapter 3 for F_a=1000c/s.

$$R \doteqdot 20 \log t - 20 \log 2\pi/\Delta F/, \qquad (2,12)$$

where $\Delta F = F_a - F_f$.

In practical condition $t_1 > t_2$ is satisfied, and t_1 is, for example, 16.6ms for the filter $/F_f - F_a/ = 33$ cps.

From (2.11) the ratio R takes a large value with time. This makes it easy to detect the formant frequency from the spectral response.

The response of bank of filters with respect to frequency, are shown in Fig.2.6 and Fig.2.7 for the formant frequency $F_f = 1000$ cps, in which the abscissa is the center frequency of filter F_a, the ordinate is amplitude or envelope of the filter response and the parameter is the time after the impulse was applied. Fig.2.6 (a) and (b) are the responses of envelopes which were redrawn from Fig.2.5(a) and (b), respectively. Fig.2.7 is the responses of amplitude redrawn from Fig.2.3. In each case, the response curves are flat for small value of t, but the formant component becomes outstanding with time. These results were also obtained from the analysis of the speech sound (see chapter 4.).

The above discussion was made for single impulse excitation. For the periodic excitation such as voiced sound, the situation is different by the harmonic structure, yielding a complicated response. However, when the fundamental frequency is low and the band widths of formant and analyzing filter are broad, the response will have the similar features to that of the single impulse excitation.

In summarizing, by using the single tuned filter bank followed by the CR smoothing network, the resultant time-frequency-intensity pattern shows different responses from that obtained by the sharp cutoff band pass filter bank followed by low pass filter type smoothing network (rather resembling to Sonagram.). In the pattern transient response is reserved, enabling the detection of the burst, the pitch pattern in voiced sound and the randomness of noise, although, for the complexity, the data to be processed for the

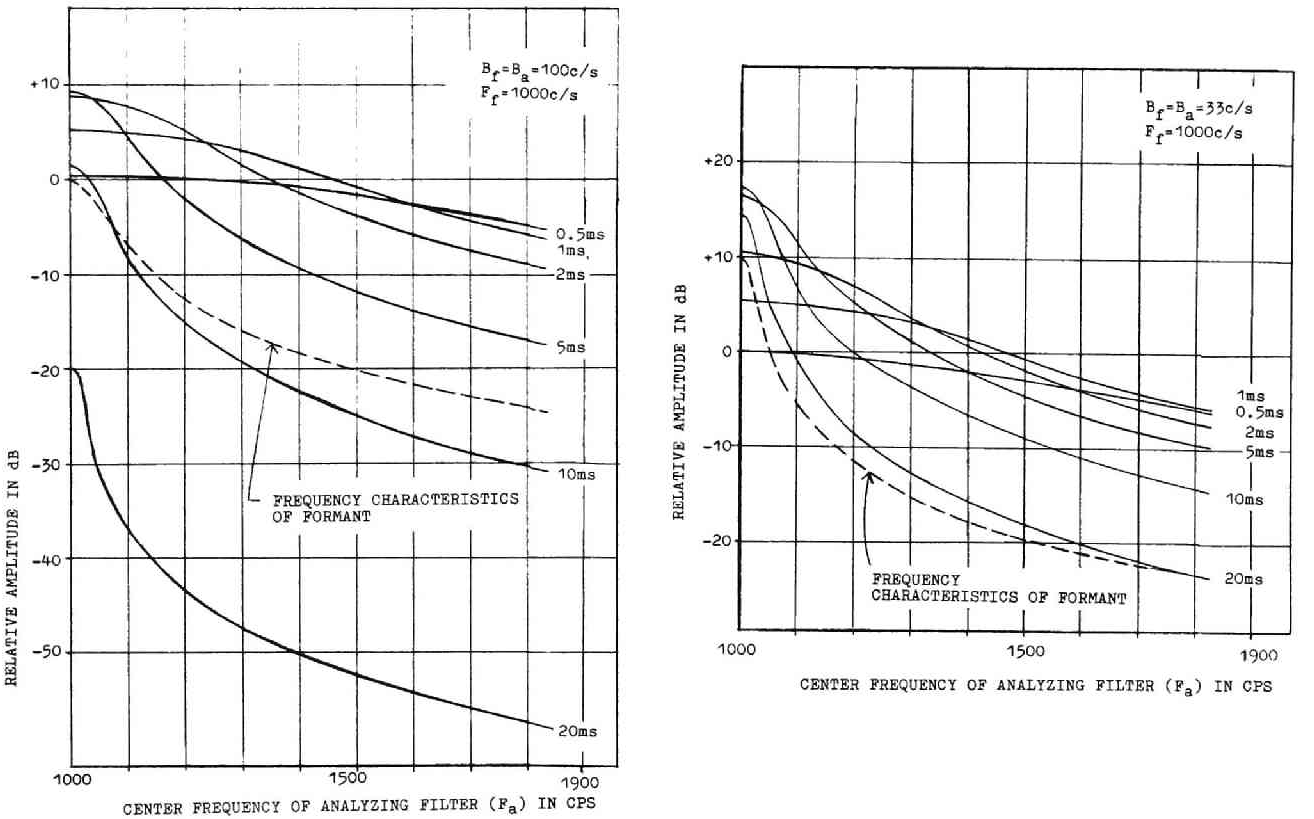


Fig.2.6 Spectral response of envelope obtained from Fig 2.5 for the condition $B_a=B_f$. Parameter is the time in milli-seconds after the impulse.

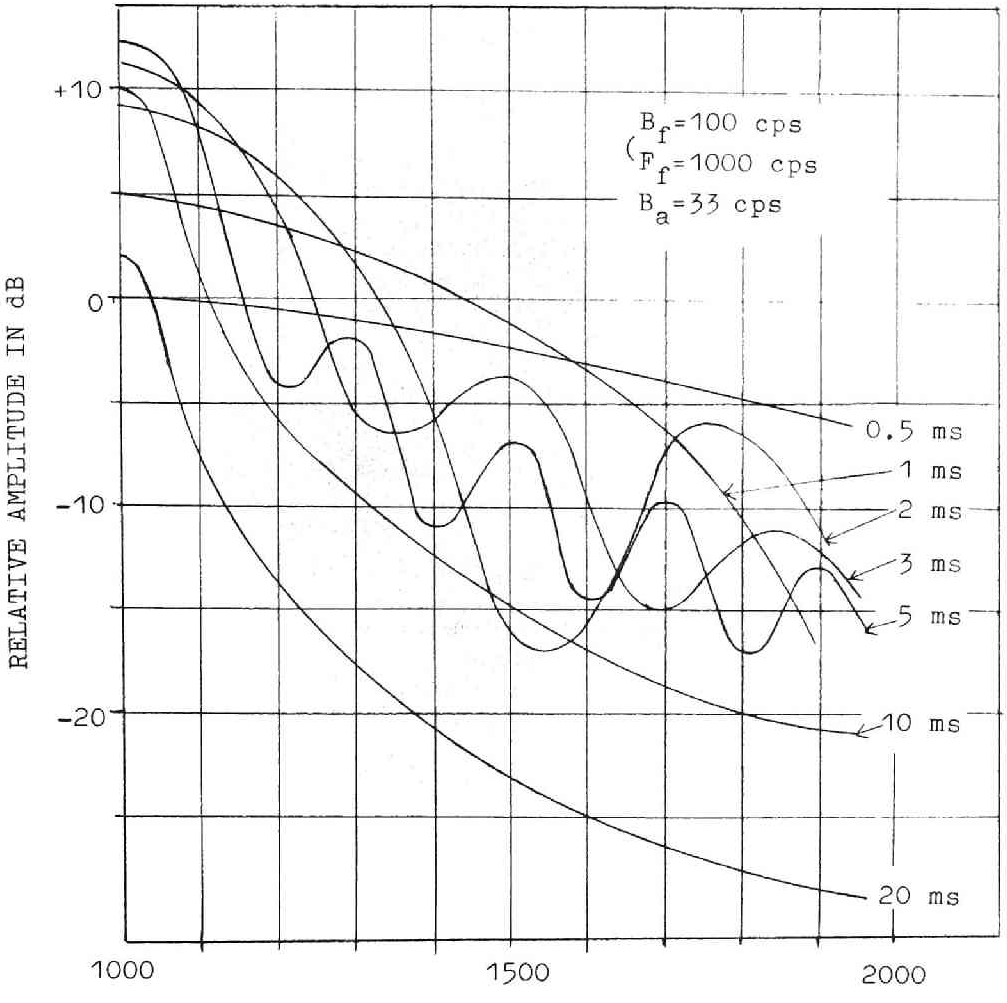


Fig. 2.7 The same as Fig. 2.6 except the condition $B_a \neq B_f$.

extraction of parameters will increase.

2.5 Conclusion

The response of a single tuned filter to a formant signal was examined. In the analysis using a single tuned filter, the time response as well as the frequency response has important information, which is a desirable feature for the analysis of transient properties of speech sound. The responses of a single tuned filter to the formant signals, having various center frequencies and excited by impulse source, were examined. The responses are characteristic to the relations of these two center frequencies and band widths. It was shown that, from the time structure of instantaneous spectrum obtained in the analysis of the formant signal by the single tuned filter bank, the component corresponding to formant frequency becomes dominant with time. The use of the condenser input type rectifier (CR smoothing network) was proposed to detect the envelope of the response in experiment. The experiment on the response of the single tuned filter bank is shown in chapter 3 and the analysis of speech sound using it is presented in chapter 4.

Chapter 3

SPECTRUM ANALYSIS BY SINGLE TUNED FILTERS

3.1 Introduction

Spectrum analysis is the most powerful means for the analysis of speech sound. Among various methods to perform it, analysis by a filter bank is popular and many analyzing devices have been designed.[2][3][8][13][17][18][42] In these devices the signal is analyzed by band pass filters and the output responses are rectified and smoothed, yielding a short-time spectrum section or spectrum pattern. The concrete constitutions are, however, different for each device. As for the analyzing filter there are several methods of realization; (i) a heterodyne method that uses one filter by sweeping its center frequency or the signal, repeating the analysis cycle several times, (ii) a bank of filter method that arranges a bank of band pass filters covering the frequency range of interest, and (iii) a combined method of (i) and (ii), which is explained below. The analyzed results, the scale of device, the object of the applications depend on these constitutions.

The spectrum analyzer presented in this chapter is semi-heterodyne type((iii) of the above classification) in which analysis is performed with repetition of several analyzing cycles, using bank of filters arranged in some frequency range within the range of interest. Single tuned filter followed by CR smoothing networks were used to be able to represent transient response in the spectrum pattern and section as stated in chapter 2. The device is also designed to be able to obtain the digital record for the computer input without degrading its quality and transient response.

The overall circuit is shown in Fig. 3.1. The speech signal is once recorded on one of the tracks of 2-track endless tape recorder with a

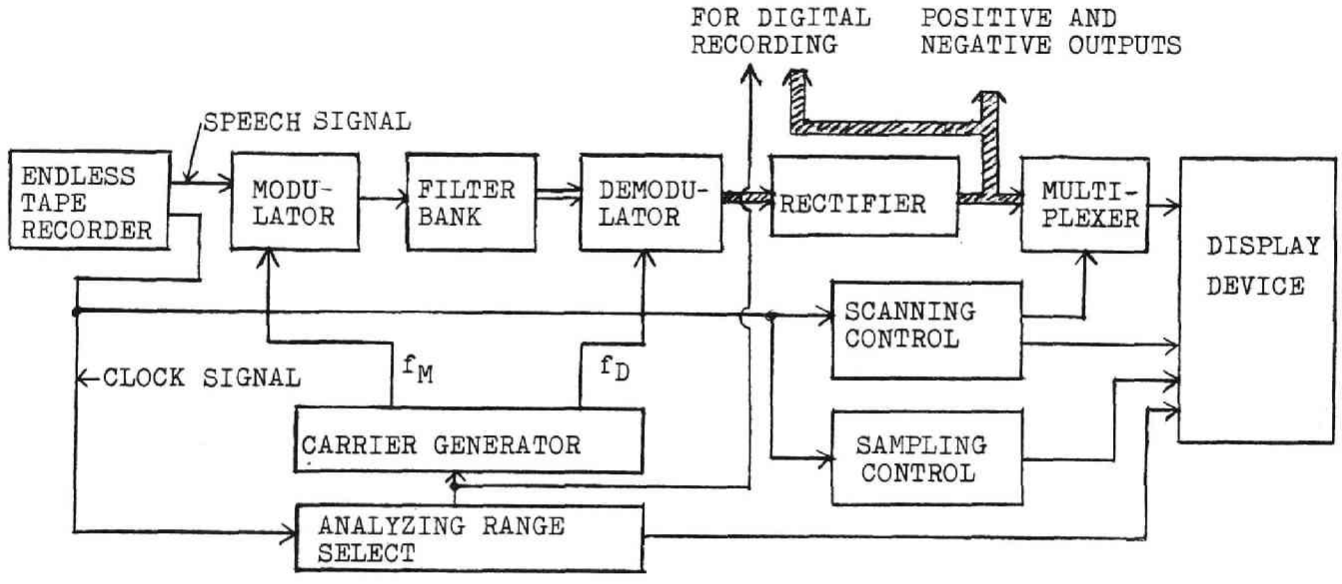


Fig. 3.1 Overall block diagram of spectrum analyzer.

repetition cycle of about 2 sec.. On recording the signal, the section of the signal to be processed is selected and the clock signals generated in that section are recorded on the other track. The speech signal to be analyzed and the gated clock signals are repeatedly reproduced, performing the analysis in each cycle. The clock is used to count the number of repetition of cycles and to hold the synchronization among the timings in each analzing cycle. All the controls of the operations are carried out based on this clock signal. The band widths of single tuned analyzing filters can be selected among 33 cps, 67 cps, 100 cps and 167 cps and the center frequency spacing between the adjacent filters is fixed to 33 cps. Therefore, when the band width of, for instance, 100 cps was used triply overupping bank of filters are formed. In such arrangement, there need 240 filters to cover 0—— 8 kc range, which may be unpractical.

One way to to solve this difficulty is to adopt the mel scale or the logarithmic scale assignment of center frequencies of filters. This is matched to the mechanism of perception of the human ear and also to the spectral distribution of information of speech. An alternative way is to use the heterodyne method, which was used here. In this case the frequency scale is essentially linear.

Fig. 3.2 shows the relation between the signal and the filter bank. 31 filters were arranged in 1 kc band at equal spacing of 33 cps. Therefore, 8 analysis cycles are needed to complete the analysis of 0 — 8 kc range of the recorded signal. For this purpose the bank is set between 13 — 14 kc and the frequency region to be analyzed is shifted up to this position by selecting the modulation carrier frequency f_M automatically. By demodulating each analyzed signal by the demodulation signal F_D with frequency f_D, the filter responses are represented in the audio range again. For constant f_D they are always in some fixed range $f_D-f_H \sim f_D-f_L$ and for $f_D=f_M$ the analyzed signal is represented in its original frequency.

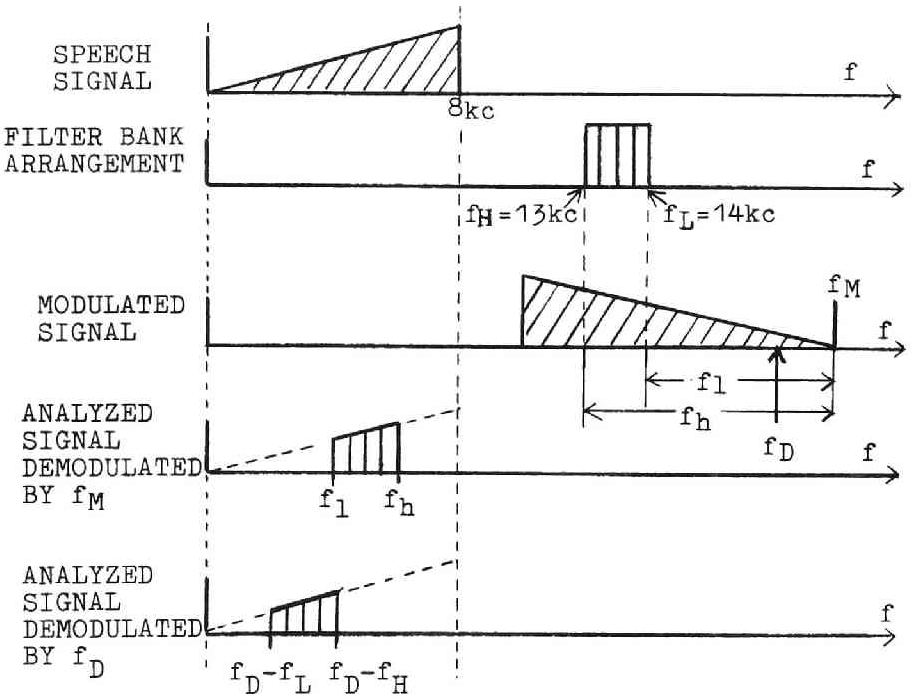


Fig.3.2 Modulation and demodulation system of signal and filter bank arrangement.

The bank of analyzers' outputs demodulated are fed to a full wave rectifier and a smoothing network in which two types of smoothing circuits are available; the low pass filter and the condenser input type CR network.

The envelopes of signals are then scanned by a multiplexer whose operation is controlled by the clock signals from the endless tape recorder. A CRT oscilloscope displays the pattern as intensity modulation and the spectrum section of selected time point, whose operations are controlled in connection with analyzer's operation. The results are recorded on photographic film.

3.2 Analyzing Device[19]

The block diagram of analyzing device is shown in Fig. 3.3. After the pre-processing the speech input is modulated by the carrier f_M which was generated by selecting some harmonic component of the 1 kc crystal oscillator. The selection is controlled manually or automatically by the carrier select signal at the end of the processing section of each analysis cycle. The filter bank is fixed in 1 kc range between 13 ~ 14 kc. By modulating the signal with modulation carrier of f_M = 14, 15, ... and 21 kc, the ranges 0 - 1 kc, 1 - 2 kc, ... and 7 - 8 kc of the input speech are analyzed, respectively. The center frequencies of filters are stabilized to ensure the equal spacing. The outputs of filters at a nigh frequency region are demodulated by demodulating carrier with the frequency f_M (the same as modulation carrier) or f_D. The selection of carrier signal is made in synchronized with the operation of the display device.

1. Preamplifier and Modulator

The input signal to the device is an audio signal which comes from endless tape recorder. After amplification the frequency range of the signal is limited up to 8 kc for the purpose to remove the components which fall in the filter arrangement. Fig. 3.4 shows the overall frequency

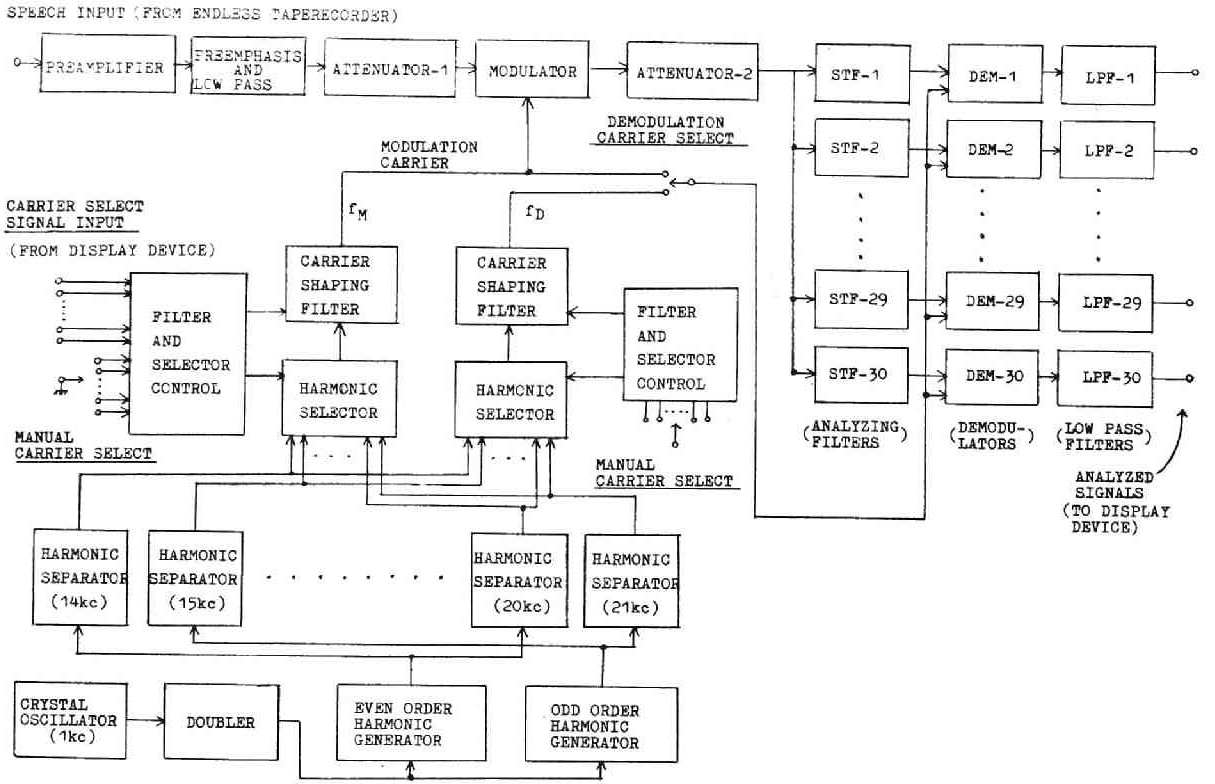


Fig. 3.3 Block diagram of analyzing device.

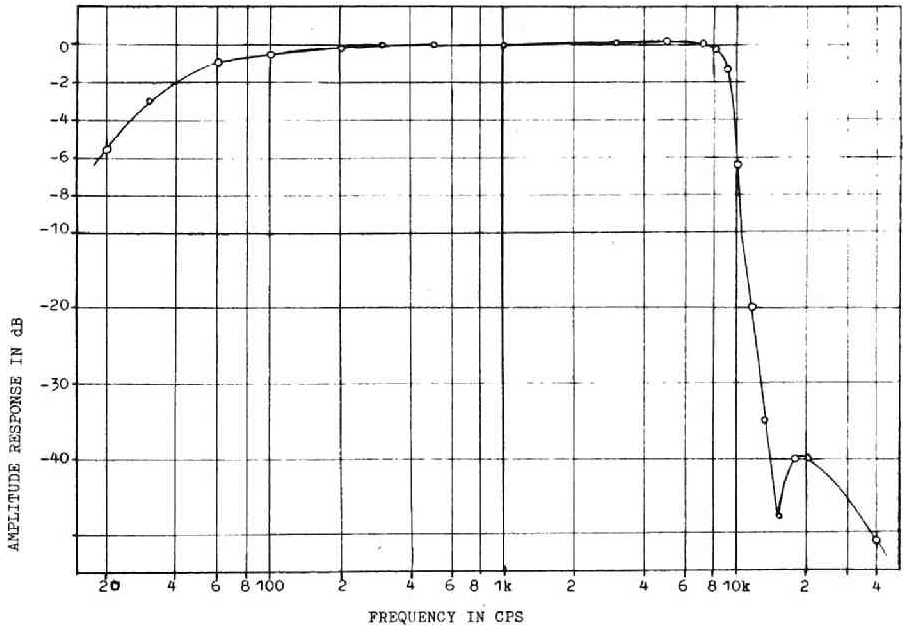


Fig. 3.4 Frequency characteristics of amplitude response of modulator output to input. Abscissa is frequency of input signal.

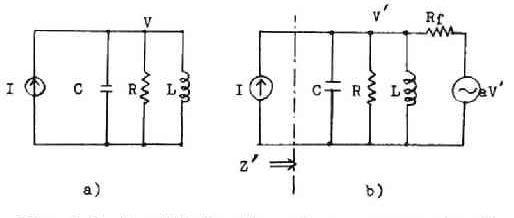


Fig. 3.5 Q multiplication of resonance circuit.

characteristics of amplitude response of modulator output to the frequency of input signal. The attenuation is gained more than 30 dB at 13～14 kc and pass band deviation is less than 1 dB in 0.2～8 kc. The attenuation in lower frequency is due to the transformer of modulator. The preemphasis circuit having time constant of 50 μs was inserted to compensate the deemphasis loss which is used in the demodulation circuit to remove the residual carrier components. The time constants of both circuits are, therefore, the same. Attenuator 1 is 12 dB attenuator in 2 dB step and attenuator 2 is 40 dB in 10 dB step. The modulator is the ring modulator by diodes, the BSB signal of which is applied to filter bank after amplification. Analysis is carried out on the lower side band components. In the case the signal frequency is very low, the other side band components will also pass the analyzing filter, For this reason the SSB modulation will be recommended.(20)

2. Generation of Carrier

The analyzing range is selected by the selection of modulation carrier frequency. The analysis is performed on each 1 kc range for each revolution of endless tape recorder, in such way as 0 — 1 kc, 1 — 2 kc, -----,7—8 kc for the carrier frequencies 14 kc, 15 kc, -----, 21 kc, respectively.

The high stability is needed in the carrier frequency because the filters are arranged in the higher frequency region. The variation of carrier causes not only the deviation of frequency scale, but the relative deviation between carrier frequencies for each analysis cycle also causes over up or drop out between the analyzing ranges. In this device, since the spacing of the center frequencies of analyzing filters is selected to 33 cps, the deviation must be less than several cycles per second, that is, less than about 0.025% for carrier frequency of 21 kc. For this requirement the carriers were obtained by selecting each of the harmonic components generated from a 1 kc crystal oscillator.

The carrier signals by this method have high stability and accuracy

with respect to frequency, but will be subjected to the amplitude variation and frequency fluttering due to the unsatisfactory suppression of sprious harmonic components. The carrier must be free from such degradations, especially for the use of demodulation.

1 kc signal from the crystal oscillator is doubled to 2 kc sine wave, which is then shaped and divided by flip-flop, yielding a rectangular signal of 1 kc having accurate duty ratio of 1/2. The odd order harmonic generator of Fig. 3.3 generates narrow pulse train of 2 kc pulse rate with alternate polarities which contain only odd order harmonics. On the other hand even order harmonic generator generates the same pulses but with one polarity having only even order harmonics. Each harmonic separator has a single tuned circuit adjusted to 14 kc, 15 kc, -----, 21 kc, respectively, which selects the harmonic component from harmonics of 2 kc spacing. The circuit has the same structure with the circuit of Fig. 3.6, but the band width is narrowed to less than 10 cps by the positive feedback, getting about 50 dB rejection of sprious harmonics.

The harmonic selector connects one of the separated harmonics to carrier shaping filter manually or controlled automatically by the signal at the end of each analysis cycle. The carrier shaping filter again shapes separated harmonic and gets pure sine wave of less than -80 dB amplitude fluttering, which is used as the modulation and the demodulation carrier.

3. Analyzing Filter

The center frequency arrangement of single tuned filter bank is as follows;

Channel No.	1	2	3	4
Center frequency (kc)	14.000	13.967	13.933	13.900
Channel No.	-----	29	30	31
Center frequency (kc)	-----	13.067	13.033	13.000

The band widths are constant for all the filters, which are selected to 33 cps, 67 cps, 100 cps and 167 cps.

The filter is composed of a parallel tuned circuit followed with a transistor amplifier. The circuit is shown in Fig. 3.6. Signal input is applied to terminal 1-1' from the modulator through attenuator -2, and the level is adjusted by RV_5. Then, it is connected to terminal No. 2 of the coil through resistor R high enough not to damp the Q value of the coil. To obtain high Q value of the resonance circuit, the feedback Q multiplication by transistor amplifier was used as well as the selection of core material. The band width is selected by SW-1 by selecting damping resistors which are adjusted to get the desired Q value. The C'_1--C'_4 are added to compensate the minor deviation of center frequencies.

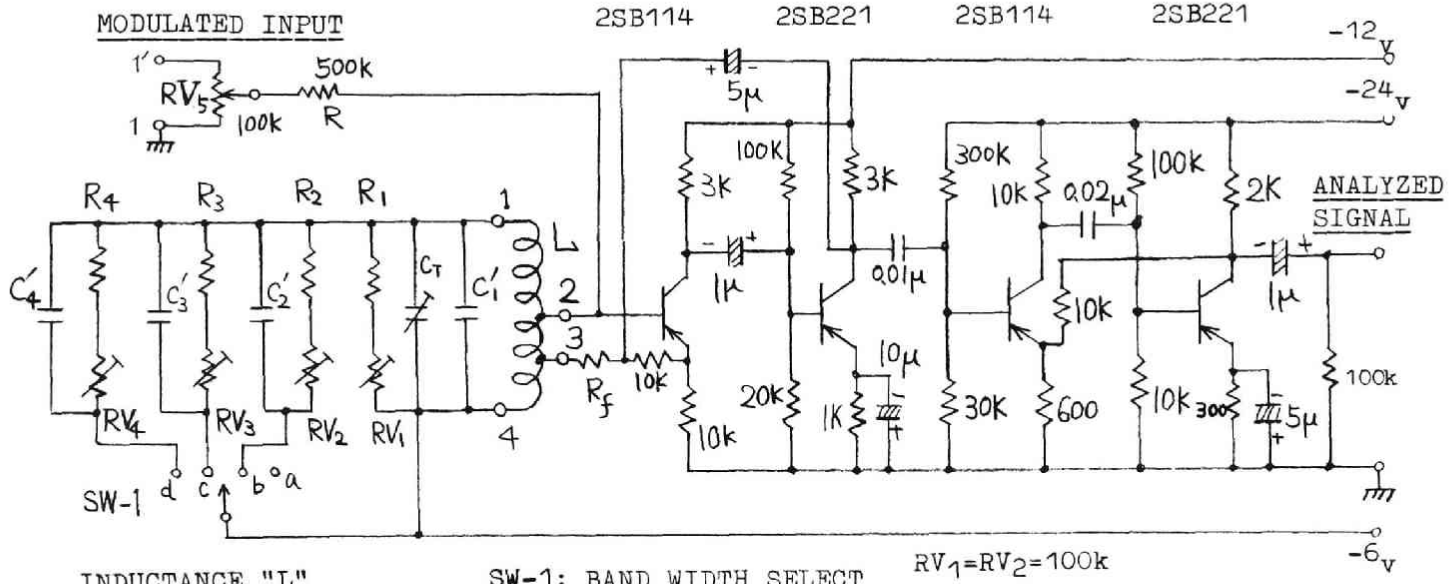
The difficult problem in this circuit is to realize a stable resonance circuit; that is, to obtain stable high Q value and to suppress the center frequency deviation by temperature. To realize 33 cps band width at, for instance, 14 kc, Q value of more than 400 is needed. If the allowed deviation is 5 cps at 14 kc for temperature shift of 20°C, then the temperature coefficient of the resonance circuit must be less than 1.8×10^{-5} /°C.

The pot type ferrite core was used as an inductor, which was combined with a styrol condenser. The positive temperature coefficient of inductor is compensated by the negative temperature coefficient of the styrol condenser (-100～-200 ppm) in such way to adjust the gap of core. The Q value thus adjusted is about 300 at 14 kc. To get it high enough, the loss of resonance circuitwas compensated by Q multiplication.

In Fig. 3.5 let's consider to boost the Q value from Q to Q'. Provided that Z is resonance impedance of parallel resonance circuit of Fig. 3.5 a), then Q is given by

$$Q = \omega CR ,$$

and at the resonance frequency $Z = V/I \quad R.$



INDUCTANCE "L"
POT CORE, μ_{eff}=175
COIL; 1-4:2-4=1:1/4
1-4:3-4=1:1/8
L≒39mH
Q≒300(AT 14kc)

C'_1: STYROL CONDENSER
C'_2, C'_3, C'_4=1——3pf

SW-1: BAND WIDTH SELECT
a: 33c/s
b: 67c/s
c: 100c/s
d: 167 c/s

R_1=1——2M
R_2≒700k
R_3≒350k
R_4≒150k

RV_1=RV_2=100k
RV_3=RV_4= 50k
R_f=70——80k

RESISTORS IN Ohm
CAPACITORS IN Farad

Fig. 3.6 Circuit diagram of analyzing filter.

Next, as shown in Fig. 3.5 b), the voltage aV' is fed back to V' through resistor R_f. In this case the impedance seen looking from the source is;

$$Z' = \frac{V'}{I} = \frac{R_f R}{R_f+(1-a)R} \equiv AR$$

The equivalent Q' in Fig. 3.5 b) is multiplied by A as follows;

$$Q' = \omega CZ' = \omega CAR = AQ$$

For a given A, necessary feedback coefficient a is given by

$$a = \frac{A-1}{A}\left(\frac{R_f}{R}\right) + 1$$

For $a \geqslant \frac{R_f}{R} + 1$, the loop will become unstable.

Now Q = 300 is obtained. Suppose the necessary Q value is Q' = 700 or band width 20 cps at 14 kc taking into account the additional damping by periferal circuits, then A = 2.3. In Fig. 3.6 the voltage of terminal No. 2 of coil is amplified by feedback amplifier with gain of about 2, whose output is returned to terminal No. 3 (the turn ratio of 3—4 winding to 2—4 is 1/2.) through feedback resistor R_f. The adjustment of Q value is made by R_f. The input impedance of transistor amplifier is the order of several hundreds kilo-ohms.

The adjustment of band width for 33 cps is made by RV_1 and the center frequency is by C_T. For the other band widths appropriate dampings are added by RV_2, RV_3 and RV_4.

4. Demodulation

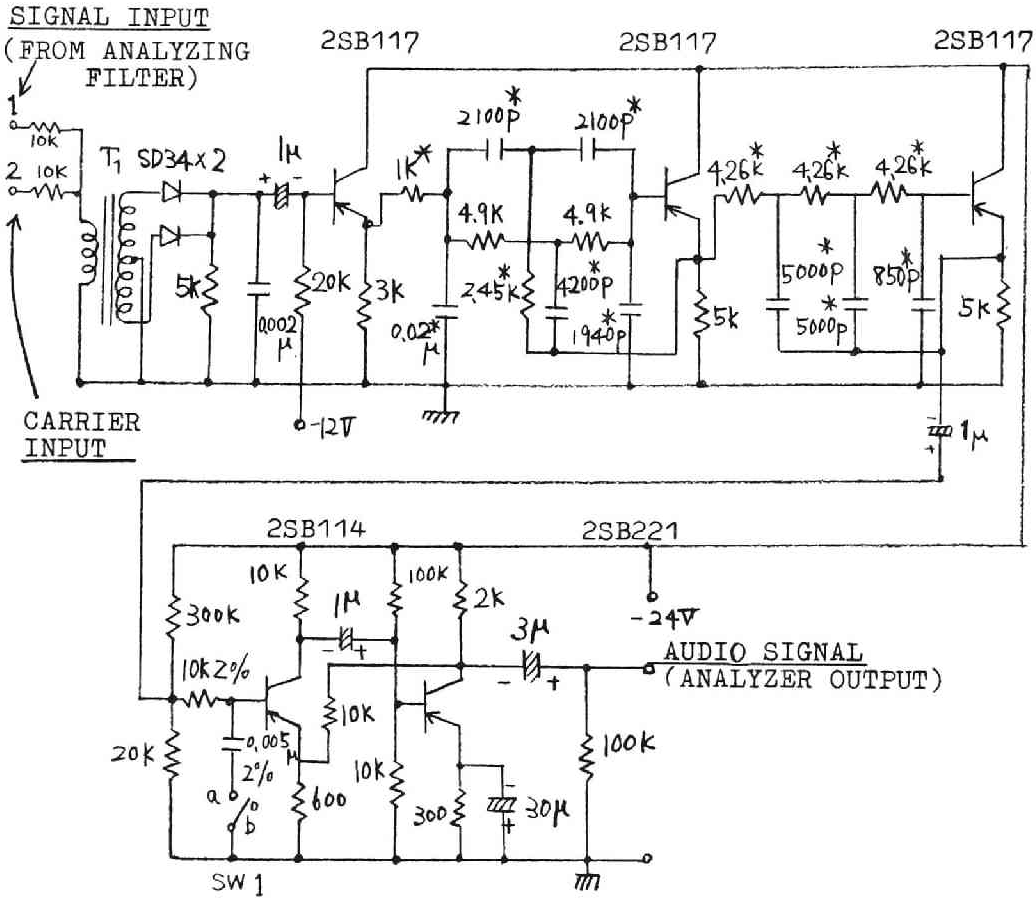
The output signals from analyzing filters are shifted by the frequency of the modulation carrier f_M from the audio signal. They may be utilized as the output of analyzer. But, to obtain the same effect as the direct filtering in audio band filters, the output of analyzing filters are demodulated. Demodulation is carried for each output and there must be prepared the same number of demodulators as analyzing filters, i.e., 31 channels. By demodulating with the same carrier as that used for the modu-

lation, the outputs are the same as the outputs of direct filtering, and on the other hand by demodulating with another carrier of constant frequency f_D, the signals are converted into an arbitrary 1 kc range within audio band (see Fig. 3.2).

The demodulation is performed by carrier injection, the circuit of which is shown in Fig. 3.7. Signal from analyzing filter is led to terminal 1 and demodulation carrier to terminal 2 which has a considerably larger amplitude than that of the input of terminal 1. After mixing these signals, resultant signal is subjected to full wave rectification, by which the carrier component included in audio signal is doubled to higher than 28 kc from about 14 kc. This makes it easy to suppress the residual carrier. The rectified signal is passed through low pass filter composed of CR active filter, which is required to have sharp cutoff characteristics. Further, to make the carrier rejection perfect, a deemphasis circuit was prepared after the low pass filter which has the time constant of 50μs, the same as that of preemphasis circuit. The overall frequency characteristics of the demodulation circuit are shown in Fig. 3.8. The pass band characteristics in (a) mean the output voltage of low pass filter when carrier of 21 kc is applied to terminal 2 and sine wave of 21 kc--12 kc to terminal 1, which shows the combined response of demodulator and low pass filter. Attenuation characteristics above 10 kc range of (b) are the frequency response of low pass filter. The resultant attenuation of carrier is more than 60 dB for 28 kc and up, together with the deemphasis circuit, which is sufficient for the analysis purpose.

3.3 Display of Spectrum[18]

A set of signals from the analyzing device is visualized as a spectrum pattern or spectrum section on cathode ray tubes. As the analyzing filter bank can not complete the analysis of full range at one time. The ana-



SW a) DEEMPHASIS(3.3kc)
b) FLAT

T_1, 2k : 1k

* ±1%

RESISTOR IN Ohm

CAPACITOR IN Farad

Fig. 3.7 Circuit diagram of demodulation circuit.

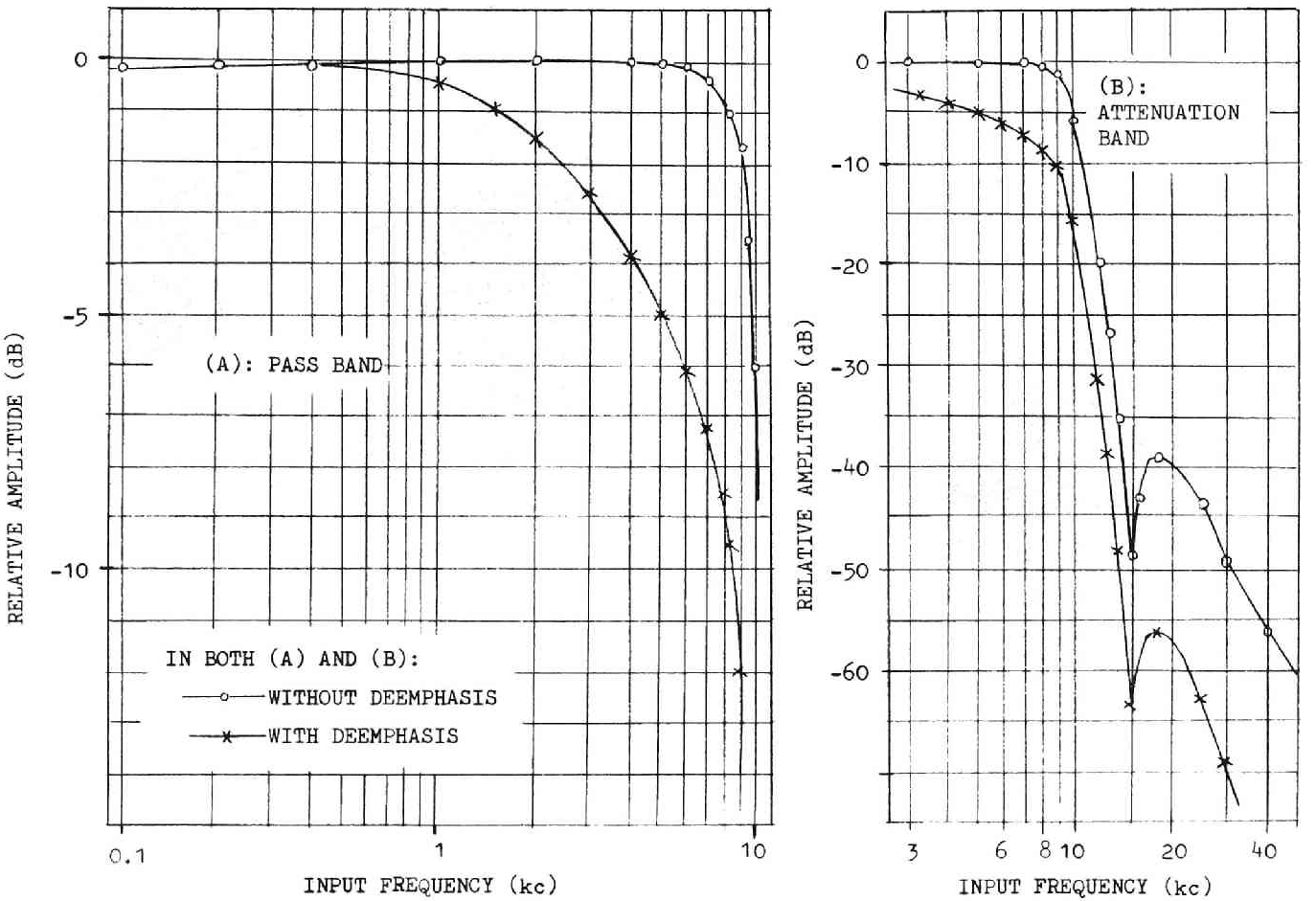


Fig.3.8 Frequency characteristics of demodulation circuit.

lysis is repeatedly performed on the signal recorded on endless tape recorder with clock pulses. The analyzed signals are amplified, rectified and smoothed and then parallel signals are converted to a serial signal, which is led to the display circuit after converted to dB scale. The block diagram of the display device is shown in Fig. 3.9.

The speech signal is transfered to track A of the endless tape recorder. (Of course the duration of signal to be analyzed must be shorter than the period of revolution of the endless tape.) At the same time the part of the speech sound to be analyzed is selected, during which the sampling clock generator produces 1 kc clock pulse and it is recorded on track B of endless tape recorder. The signal on both tracks are synchronously reproduced several times. The speech signal is led to the analyzing device and the clock pulse is to the other part of the display device to control all the operations. By doubling the reproduction speed, the effective band width of 16.5 cps may be obtained and, by halving it, the normal sampling period of 2 ms can be effectively reduced to 1 ms. By detecting the envelope of clock pulses the beginning and the ending point of the analysis section of each reproduction cycle are obtained from signal A. The beginning signal starts the display circuit and the ending signal steps up the analyzing range counter by one.

The clock pulse is used to control the multiplexer scanning. For each clock, the astable multivibrator (A. M.) control flip-flop is set, by which scan signal generator starts its oscillation of 20 kc, the phase of which being synchronized with the clock pulse. The 20 kc pulses are counted till it reaches the value preset by the scanning channel number selector, after which A. M. control flip-flop and counters are reset. The sequential pulse signals P_0, P_1, ·····, P_{30} to drive the corresponding switch of the multiplexer are decoded from counter outputs. The channel number is set by the scan channel number selector.

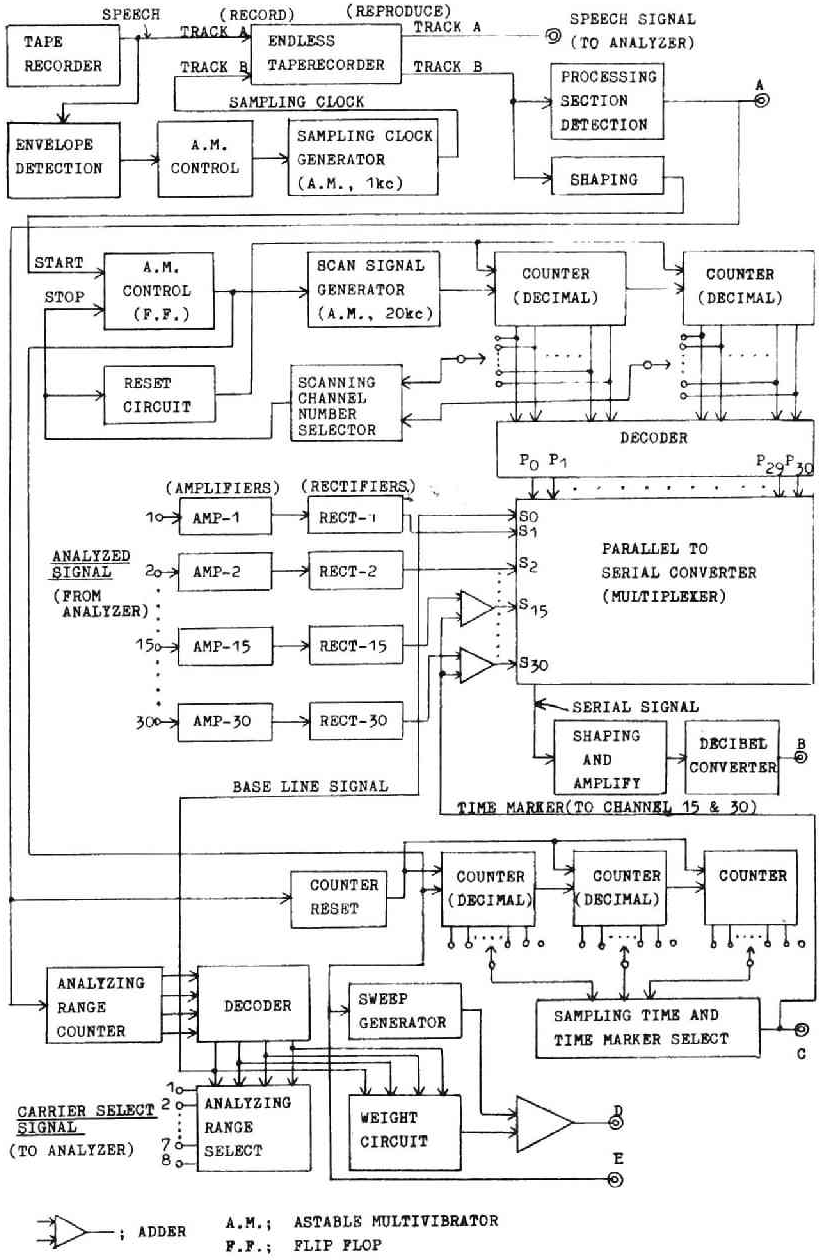


Fig. 3.9 Block diagram of display device.

The multiplexer is composed of analog switches which converts parallel signals from rectifiers to a serial signal. The switch elements is shown in Fig. 3.10. The analog signals of positive polarity S_i (i=1,2,-----, 30) are switched by negative pulses P_i (i=1,2,-----,30), respectively, and switched outputs are gathered to common load R. The sampling speed is 20 kc and the dynamic range is more than 40 dB, requiring 1.5 ms for one sampling field.

The analyzed signals from the analyzer (30 channels) are amplified, rectified and smoothed by the circuit of Fig. 3.11, by which envelopes of the signals are detected. The audio signal is amplified and rectified by the bridged rectifier. The diode pair in feedback loop compensates the rectification loss at low level. The level characteristics are shown in Fig. 3.12 which has the dynamic range of more than 40 dB.

The output of the rectifier is smoothed to detect the envelope. Two types of circuits are prepared which are selected by switch SW-1 of Fig. 3.11. One is the ordinary low pass filter for detecting the averaged instantaneous power of spectrum, the characteristics of which are designed so as to be able to smooth out the pitch synchronized fluctuation of the envelope as shown in Fig. 3.13. The alternative type is CR network combined with the rectification circuit, or the condenser input type rectifier, having fast charging time constant less than 0.1 ms and slow discharging time constant of 1 ms or 3 ms. These values were selected to follow the envelope of filter response as has been stated in chapter 2. The negative output as well as the positive polarity output is available from terminal OUT-.

The dB representation may be desired for the spectrum section representation, and also for the spectrum pattern representation to compress the signal dynamic range. For this purpose the deci-Bel converter of Fig. 3.14 is connected to the multiplexer output after amplification. The circuit performs linear approximation at the steps of 6 dB for input level

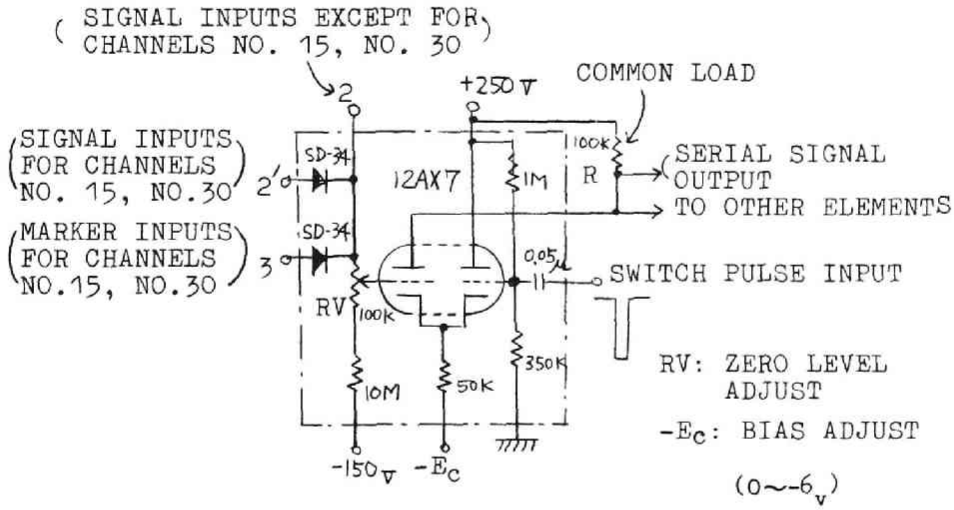


Fig 3.10 Analog switch element of multiplexer.

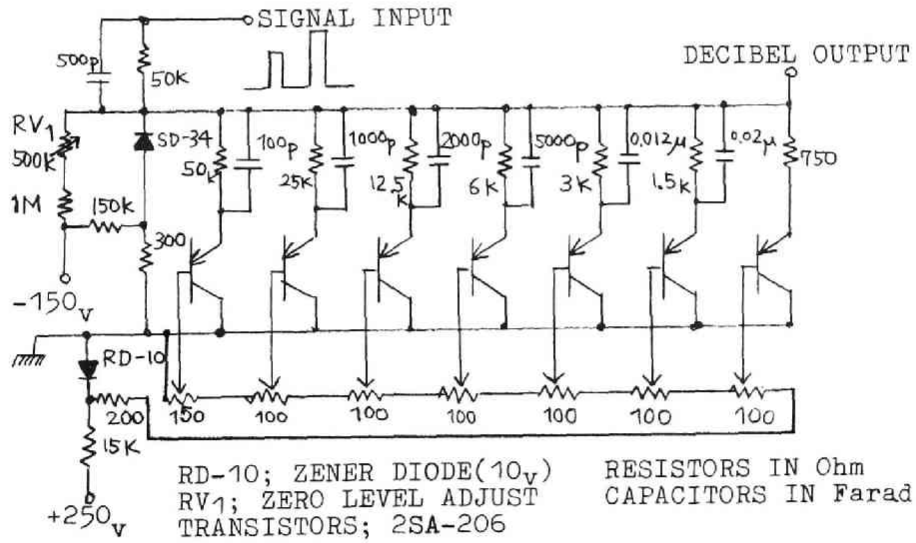
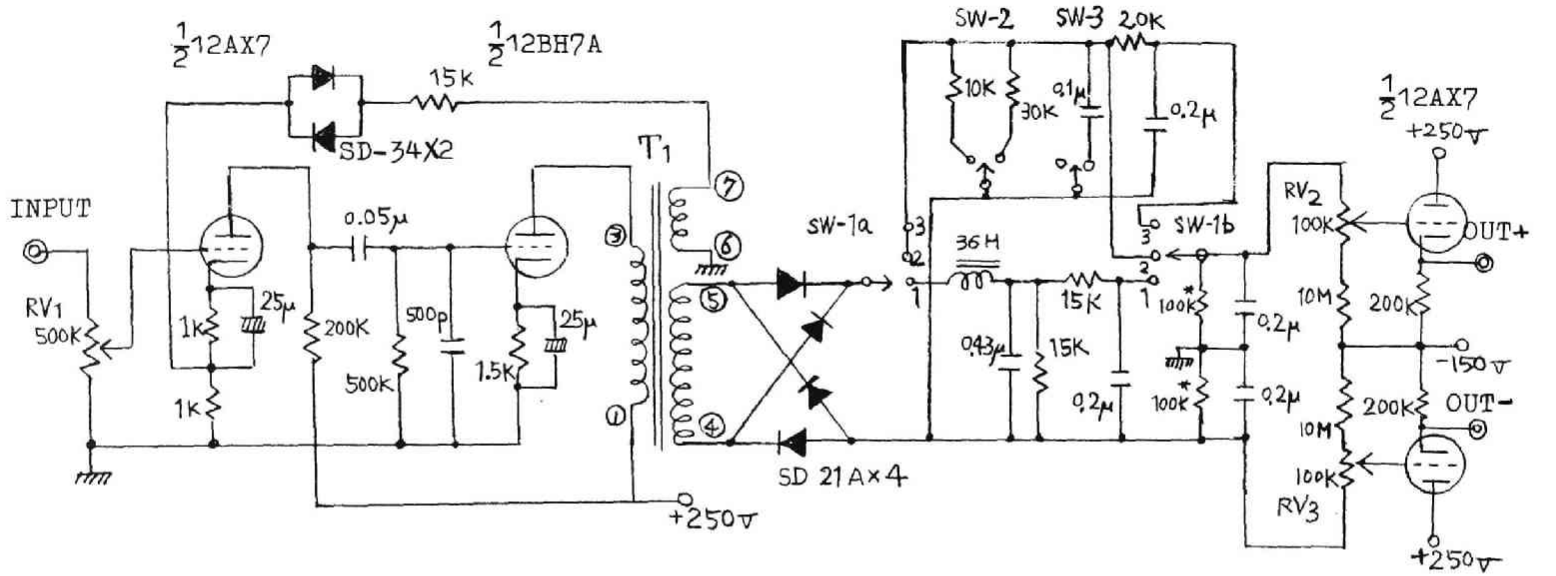


Fig. 3.14 Deci-Bel converter.



T_1; ①-③:④-⑤ = 3:1
①-③:⑥-⑦ = 30:1

SW_1; SMOOTHING CIRCUIT SELECT
1; LOW PASS FILTER
2; CONDENSER INPUT TYPE
3; CONDENSER INPUT TYPE
PLUS CR LOW PASS FILTER

SW-2 &SW-3; TIME CONSTANT SELECT
RV_2 &RV_3; ZERO LEVEL ADJUST
* ; ±1%

Fig. 3.11 Circuit diagram of amplifier and rectifier with low level compensation and smoothing network.

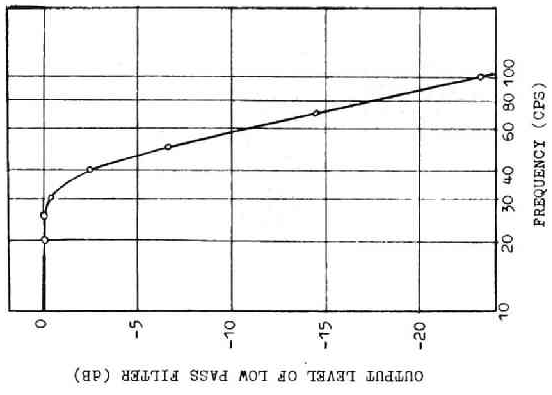


Fig. 3.13 Frequency characteristics of smoothing low pass filter.

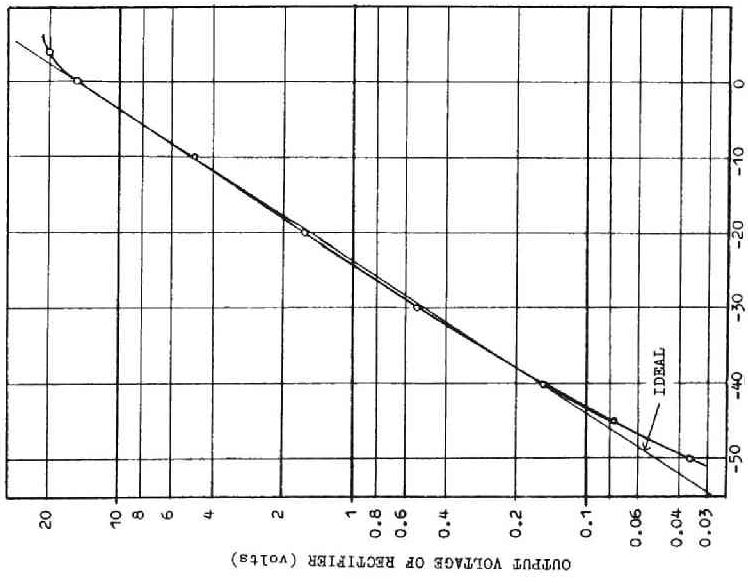


Fig. 3.12 Typical characteristics of amplifier and rectifier with low level compensation.

1 volt—100 volts, giving the output of 1 volt/6 dB.

Fig. 3.15 shows the overall level characteristics of amplifier, rectifier, multiplexer and dB converter, i.e., the relation of output voltage of deci-Bel converter to the input level of amplifier.

The results are visualized on a picture tube of the pattern display device and on a cathode ray tube of synchroscope as spectrum pattern and spectrum section, respectively. Spectrum pattern is displayed on the picture tube with long persistence phosphers by intensity modulation. The horizontal axis is time which starts the sweep at the beginning point of signal A of Fig. 3.9. The vertical axis, the frequency axis, is deflected by the signal D of Fig. 3.9 which is the summation of the sweep signal synchronized with the multiplexer scanning and the output of weight circuit which increases by constant value in connection with the step up of the analyzing range. Marker signal is provided as bright spots normally in 500 cps spacing, and in time interval of 10 ms, 20 ms, 100 ms, etc., which can be selected by the sampling time and time marker select circuit.

Spectrum section is displayed on the cathode ray tube of synchroscope. The sampling point is set at the sampling time select circuit, by selecting the specified scanning start signal of multiplexer. The synchronization of each cycle is held by the clock pulse. The connections of output signals A, B, C, D and E of Fig. 3.9 to display devices are given in Table 3.1.

3.4 Response of the Spectrum Analyzer

In spectrum analyzer stated above the filter bank is composed of single tuned band pass filters having the band widths of 33 cps, 67 cps, 100 cps and and 167 cps and the linear center frequency arrangement of 33 cps spacing. As the smoothing circuit after the rectification, the LC network (ordinary low pass filter) and the CR network (condenser input type rectifier) are

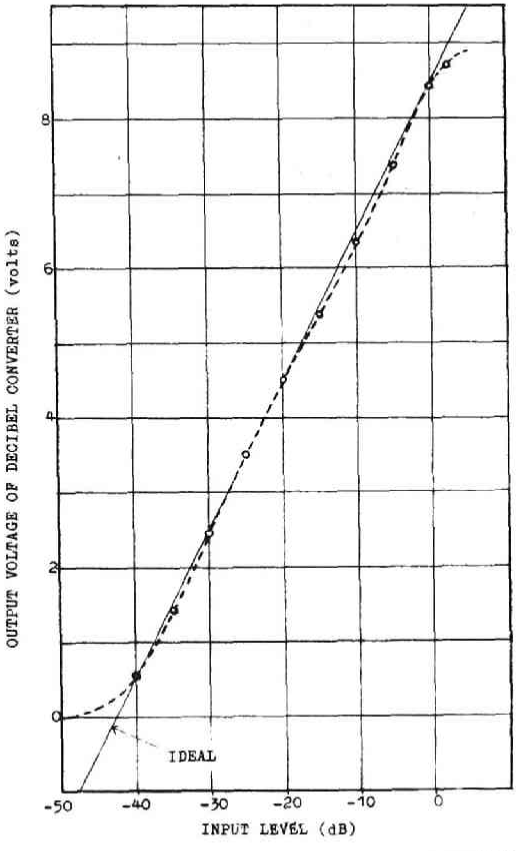


Fig. 3.15 Overall characteristics of amplifier, rectifier, multiplexer and deci-Bel converter.

Table 3.1 Connection of signals for the control of display devices. Signals (A),(B),(C),(D)and (E) are given in Fig. 3.9.

	SPECTRUM PATTERN		SPECTRUM SECTION	
	PARAMETER	SIGNAL	PARAMETER	SIGNAL
HORIZONTAL AXIS	TIME	INTERNAL SWEEP TRIGGERED BY (A)	FREQUENCY	EXTERNAL SWEEP BY (D)
VERTICAL AXIS	FREQUENCY	EXTERNAL SWEEP BY (D) UNBLANKING BY(E)	AMPLITUDE OF SPECTRUM	(B)
INTENSITY AXIS	AMPLITUDE OF SPECTRUM	INTENSITY MODULATION BY (B)	SAMPLING TIME POINT	INTENSITY MODULATION BY (C)

prepared which give different analyzed patterns. The results are observed on cathode ray tubes as time-frequency-intensity spectrum pattern (pattern) and as frequency-amplitude pattern (section) at a sampled time point, which were photographed on film.

Before the analysis is made on the speech sound, the responses of the analyzing filter bank to the various types of signals were examined for the different conditions such as band width, smoothing network, etc.. The input signals used are; sine wave, white noise, impulse train of several fundamental frequencies and formant shaped signal. The results are shown in Figs. 3.16—3.26. To read these results, please refer to "Notes on reading photographic data" in APPENDIX I of the supplements.

(i) Spectral response of analyzer: Fig. 3.16(a) is the level characteristics where the input level is shifted by 10 dB step for the filter band width of 33 cps. The dynamic range is about 40 dB.

(ii) Response of filter bank: Fig. 3.16(b) is the responses of the filter bank with different band widths to a sine wave of 1 kc.

(iii) Response to white noise: White noise used is generated by noise diode. In Fig. 3.17, the spectrum sections (No. 1, No. 2) are obtained by using the LC network, and the patterns are taken for several combinations of band widths and smoothing networks. In all cases the randomness of white noise is reserved, but the detailed appearance seems to depend on the processing method. The response to the white noise shaped by 2 kc resonance circuit having the band width of 100 cps is shown in No. 8.

(iv) Impulse train: In Fig. 3.18, the input is a wide band impulse train with constant fundamental frequency of F_p=100 cps. For the narrow band filter, horizontal harmonic bars are observed, while for the wide band the vertical lines are seen. Fig. 3.19 (A), (B) shows how the harmonic structure is formed with time in the spectrum section for the case of LC smoothing. For the higher pitch and the narrower band width, harmonic separation is

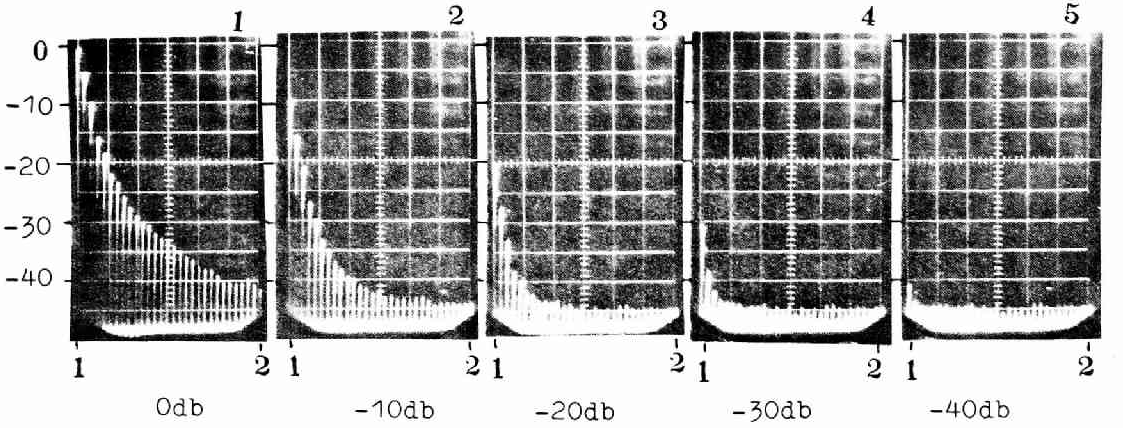


Fig 3.16(a) Level characteristics of spectrum analyzer. Input level is shifted every 10 dB. Abscissa; center frequency of filters (kc), Ordinate; output level(dB) Input; sine wave of 1 kc. Filter band width; 33 c/s

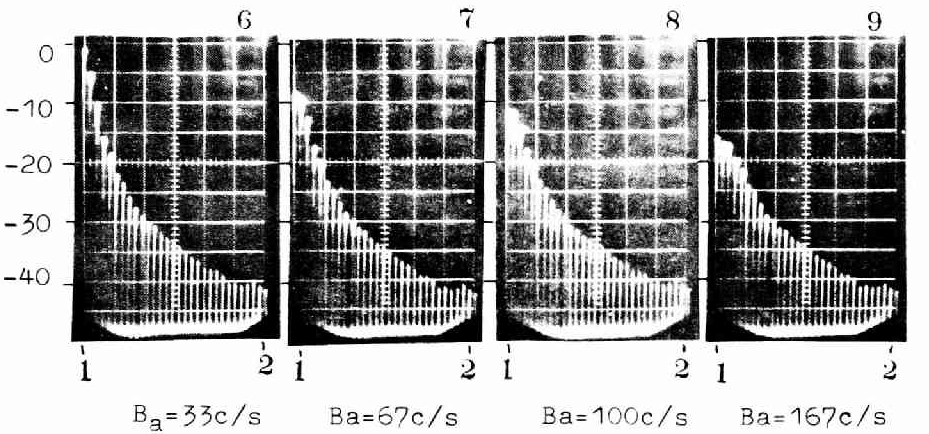


Fig. 3.16(b) Spectral response of the bank of single tuned filter with different band widths to sine wave of 1 kc.
Abscissa; frequency (kc)
Ordinate; output level(dB).

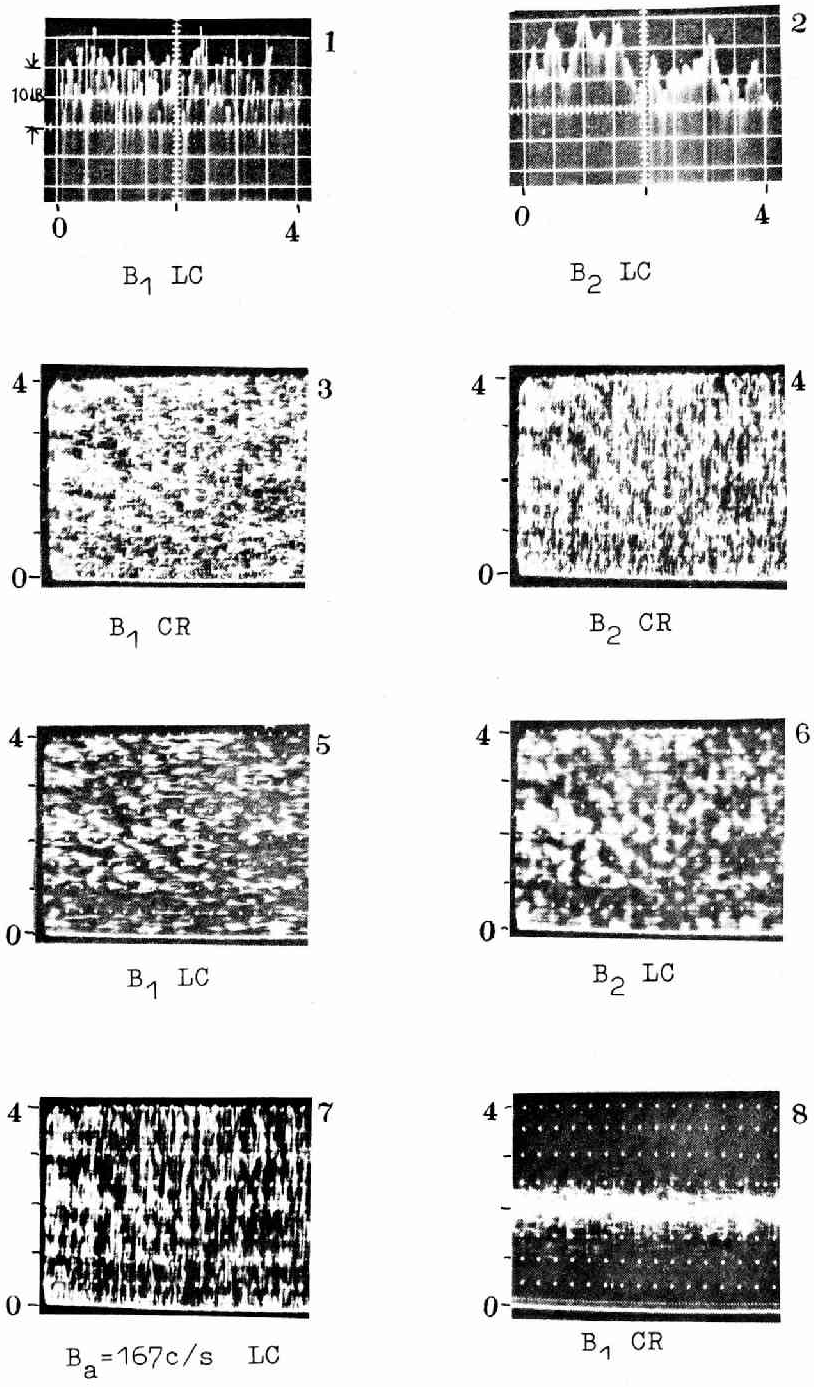
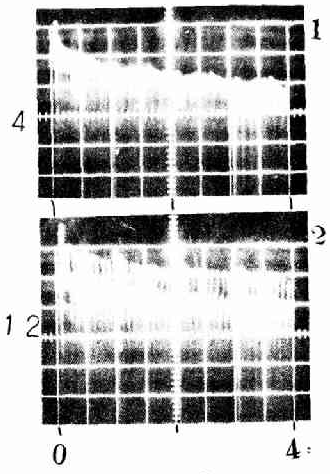
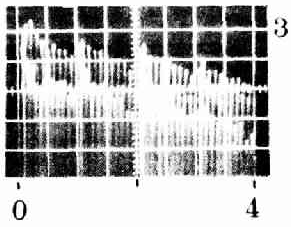


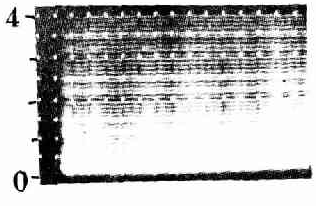
Fig. 3.17 Response of single tuned filter bank to white noise.



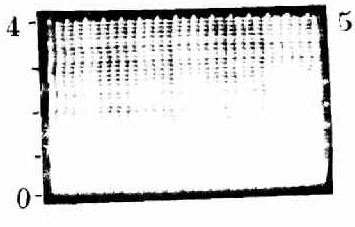
B1 LC
(t=0 at the first exitation.)



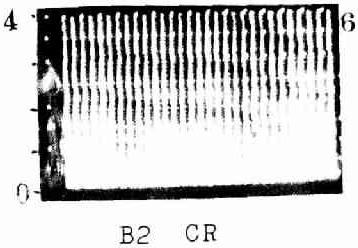
B1 CR



B1 LC

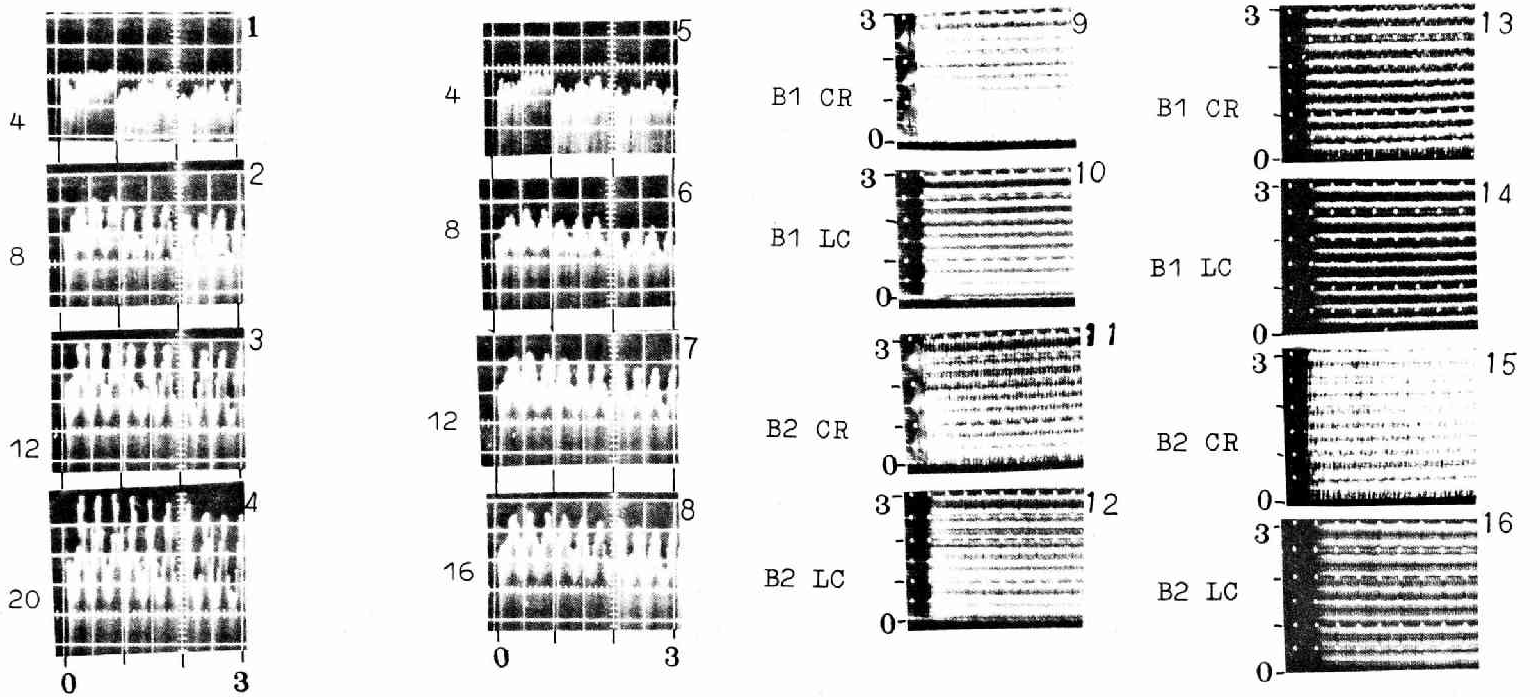


B1 CR



B2 CR

Fig. 3.18 Response to the impulse train of F_p=100c/s with unit step envelope.



(A) B1,LC (B) B2,LC (C) (D)

Fig. 3.19 Response of the bank of single tuned filter to the impulse train of F_p=300c/s. (A),(B),and (C) ; with unit step envelope. (D); with envelope of 20ms rise time constant.

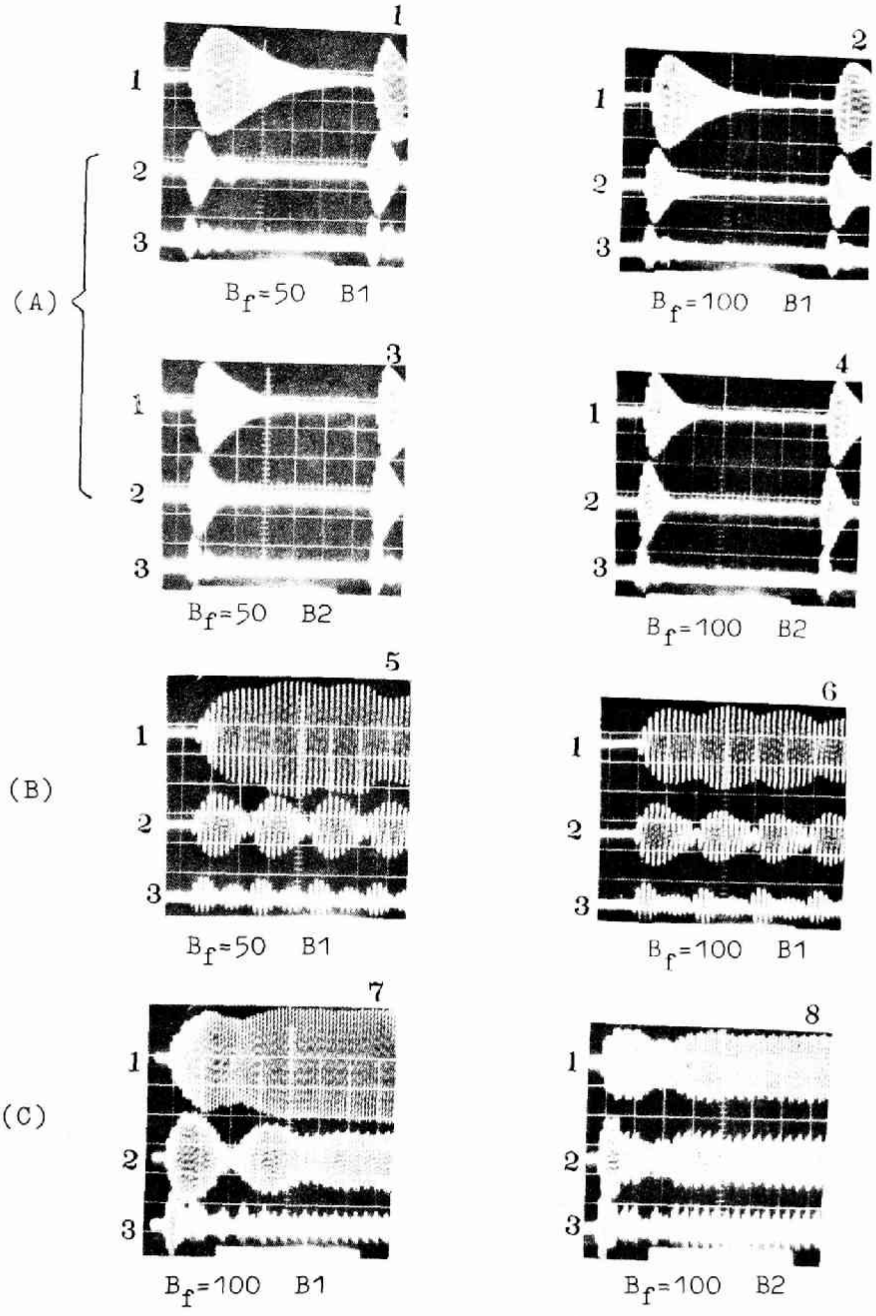
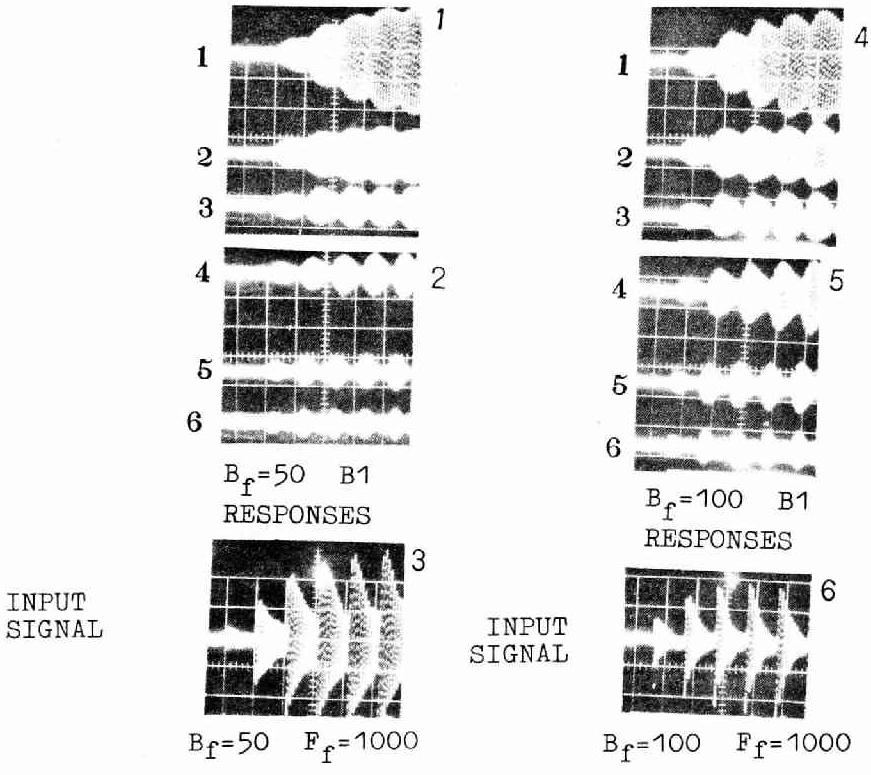
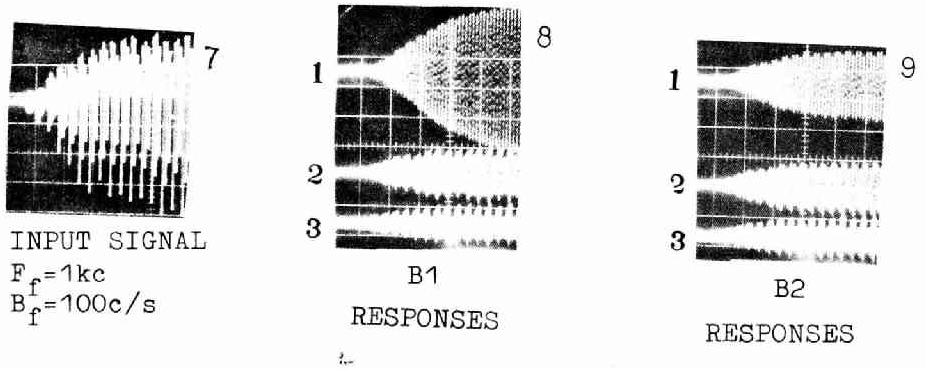


Fig. 3.20 Response of filters to formant(F_f=1kc). Excitation frequency of formant circuit is F_p=100c/s in (B) and F_p=300c/s in (C). Abscissa; 10ms/div. Channel NO. 1; F_a=1000c/s, NO.2; F_a=1033c/s NO.3; F_a=1067c/s.



(A) F_p=100c/s
Channel NO.1; F_a=1000c/s, NO.2; F_a=1033c/s, NO.3; F_a=1067c/s, NO.4; F_a=1100c/s, NO.5; F_a=1133c/s NO.6; F_a=1167c/s.



(B) F_p=300c/s. Channel NO.1; F_a=900c/s, NO.2; F_a=966c/s, NO.3; F_a=1033c/s.

Fig. 3.21 Response of filters to formant(Ff=1000c/s). Formant circuit was excited by the impulse train with envelope rise time constant of 20ms. Abscissa is 10ms/div.

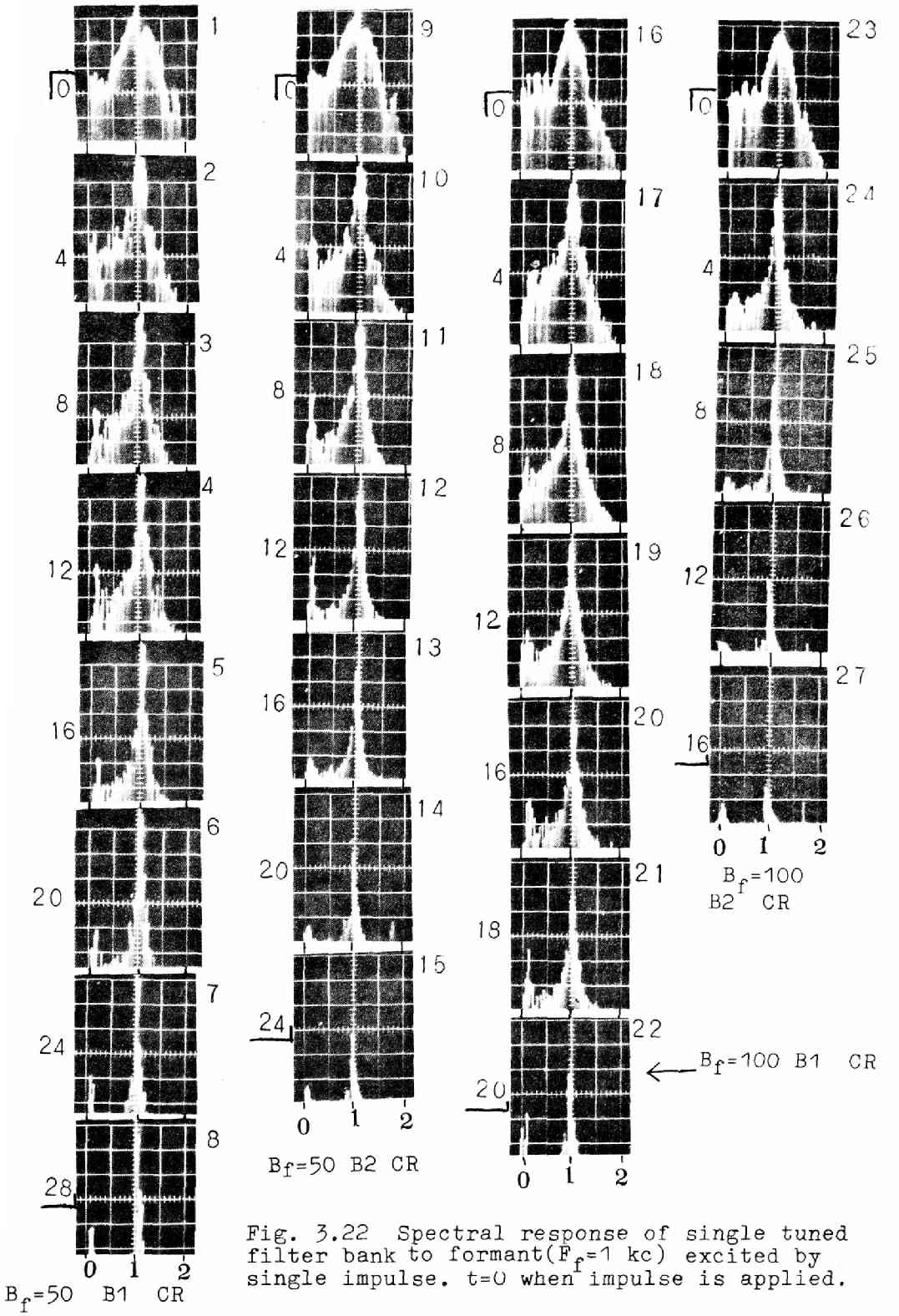


Fig. 3.22 Spectral response of single tuned filter bank to formant(F_f=1 kc) excited by single impulse. t=0 when impulse is applied.

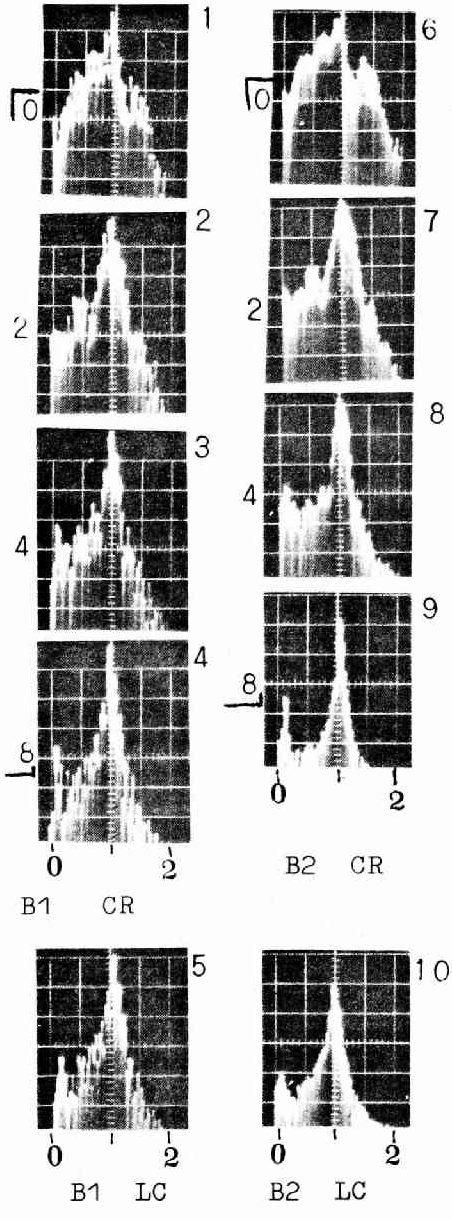


Fig. 3.23 Steady state spectral response of single tuned filter bank to formant excited by impulse train. F_p=100c/s, B_f=100c/s. t=0 when the formant circuit is excited by an impulse.

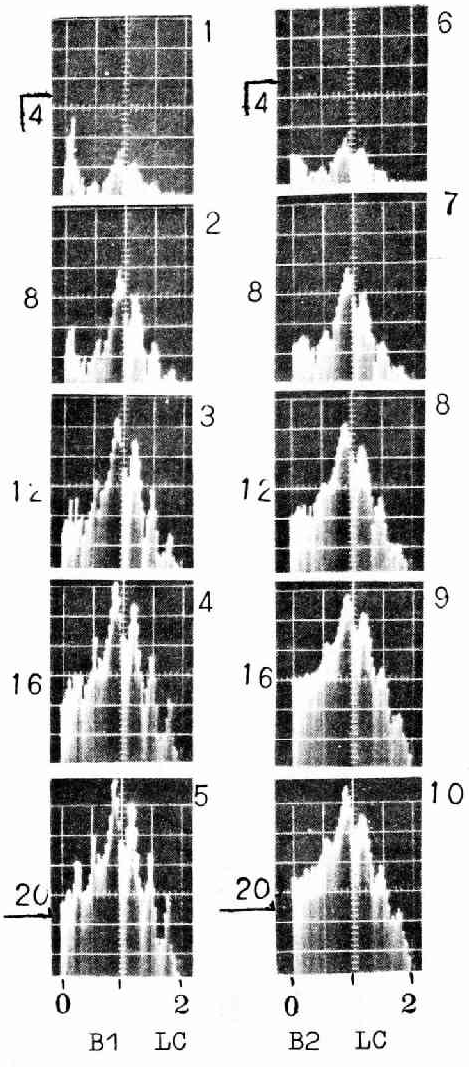


Fig. 3.25 The same as Fig. 3.24 except the condition F_p=300c/s.

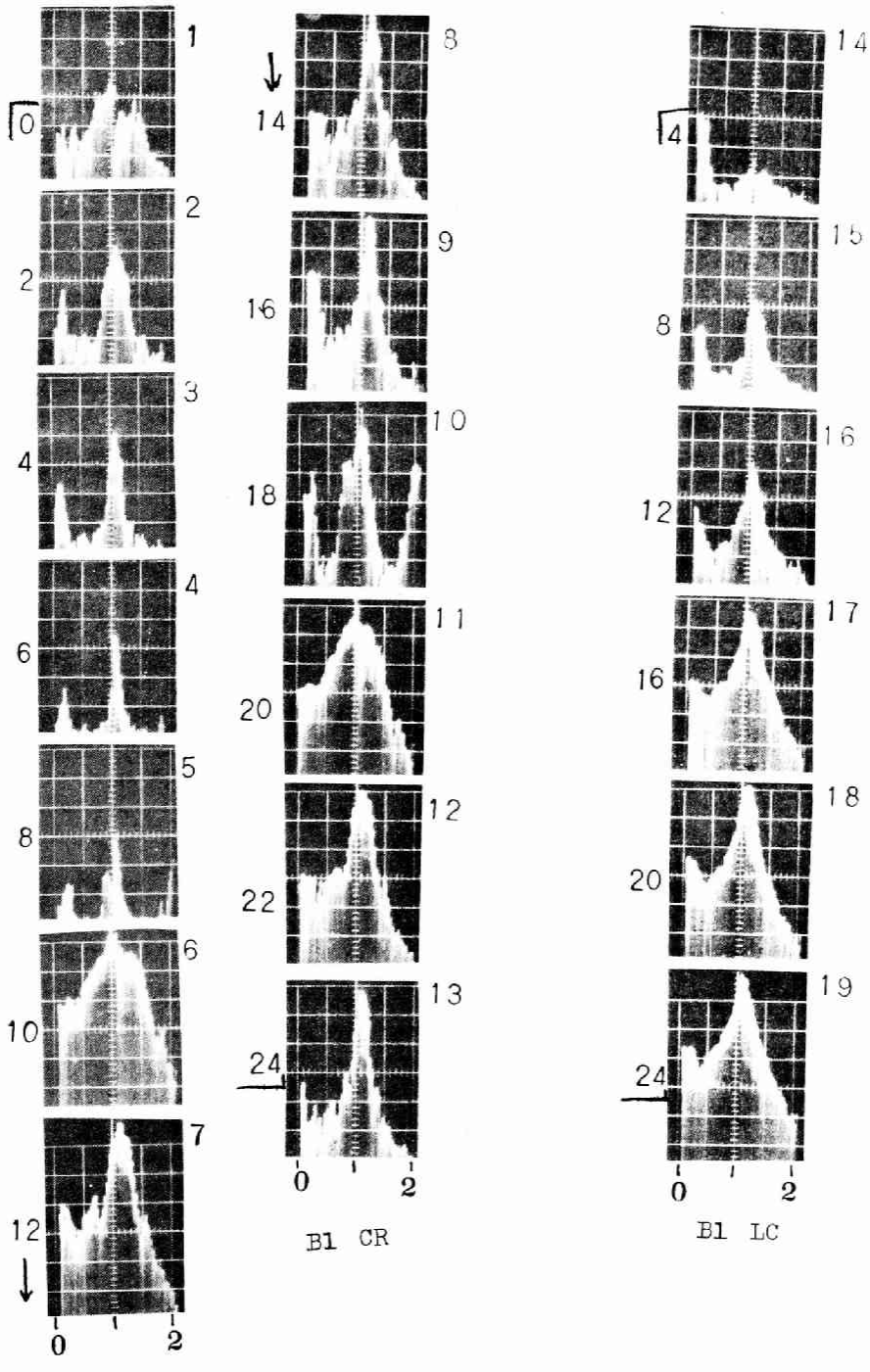


Fig. 3.24 Spectral response of single tuned filter bank to formant of F_f=1kc and B_f=100c/s, excited by impulse train with envelope of 20ms rise time constant. F_p=100c/s. t=0 at the first excitation of impulse.

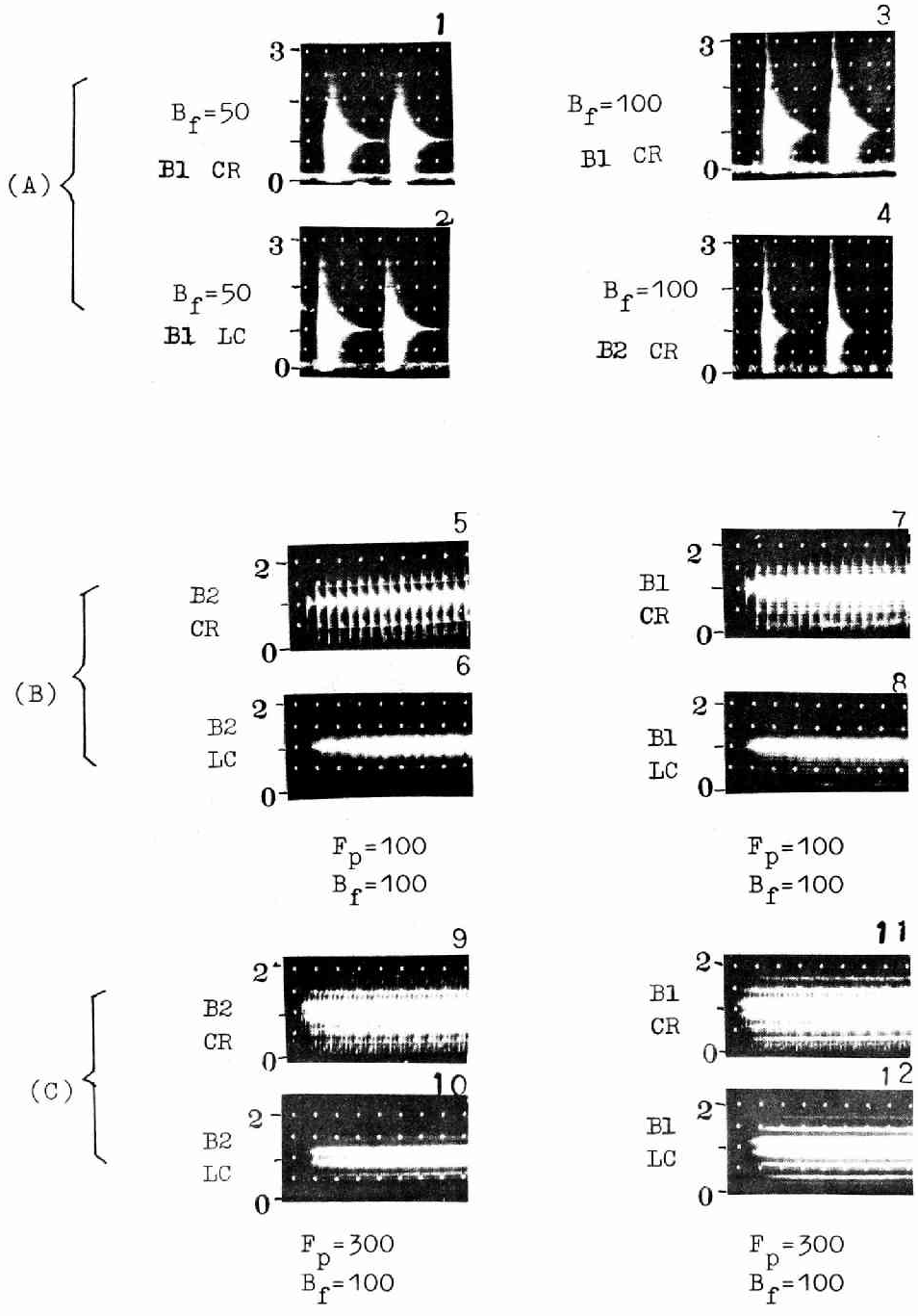


Fig. 3.26 Spectrum pattern to formant of F_f=1000c/s for different fundamental frequencies.

sufficient. (C) and (D) are the harmonic patterns for the impulse train with a unit step envelope and with a envelope of 20 ms rise time constant, respectively.

(v) Response to formant: Fig. 3.20 (A) is the time response of the filter output to the formant signal of F_f=1 kc excited by impulse of low repetition rate, which will be compared with the results by the calculation of chapter 2. The response for a formant driven by the gated impulse train of F_p=100 cps with unit step envelope is shown in Fig. 3.20 (B) and for F_p=300 cps in (C). The appearances of initial part of these cases are fairly different from those obtained for the impulse train excitation with a certain envelope rise time as shown in Fig. 3.21 (A) and (B), which is also the case with the speech signal. The change of spectrum sections with time after the impulse was applied to a formant circuit is shown in Fig. 3.22 under various conditions. The sharpening of pattern with time is observed as expected from Fig. 2.6. For the higher fundamental frequency, the situation becomes complicated but the change of spectrum shape synchronized to the excitation of the formant circuit is observed as shown in Fig. 3.23. Also, Fig. 3.24 and Fig. 3.25 show the spectrum change at the onset of a formant signal driven by an impulse train having envelope of some rise time constant.

Fig. 3.26 is the pattern representation for a formant signal. In (A) excitation is made by very low frequency in which the sharpening effect is visualized. For the lower fundamental frequency (with wide band filter and CR smoothing network) a pitch synchronized pattern is clearly obtained, while for the higher fundamental frequency the separated harmonic bars are dominant.

The above results for several specified signals may be useful to estimate the responses for the actual speech signal.

3.5 Conclusion

The spectrum analyzer using a single tuned filter bank and its responses to various types of signals were described in this chapter. By using carrier generated by a crystal oscillator and by canselling the temperature coefficients of inductance and capacitance of the resonance circuit, the stability of center frequency of each filter was sufficient for the analysis of speech sound. The dynamic range of the analyzed signal demodulated to audio band is more than 60 dB. The over all range of display circuits is more than 40 dB, when the envelope is expressed in dB scale. As stated in chapter 2, after the rectification, CR smoothing networks having various decay time constants were used as well as LC low pass network. By repeating analysis procedure 8 times under the control of clock pulses recorded on the magnetic tape loop with signal, 240 filters were effectively realized in 0 - 8 kc range. The device was designed so as to be able to introduce all the results observed by display devices into the computer. The responses to the various types of signals showed the characteristic patterns, which serve to read the analyzed results of the speech sound as shown in chapter 4.

Chapter 4

SPECTRUM ANALYSIS OF JAPANESE SPEECH SOUNDS

4.1 Introduction

Much effort has been made on the spectrum analysis of the speech sound, especially using Sonagraph. On the other hand, several kinds of bank of filter type analyzers were developed and thereby analyses were tried. The results obtained by these methods depend on the characteristics and the constitution of the particular system. In general Sonagraph can perform detailed analysis in place of the time requirement, while the bank of filter type makes a simplified representation in short time.

In most cases the smoothed power spectrum was utilized by smoothing the outputs of filters. As a rule these methods may be useful for the analyses of stationary or quasi-stationary parts of the speech sound which occupy large percentage of the speech sound. The transient part is very short in time, but it is important especially in consonant sounds, for which the processing from the time response point of view must be considered. In this chapter analyses of speech sounds were carried out in taking into consideration the time response as well as the power spectrum.[19]

1. Experiment Procedure.

The spectrum analyzer described in chapter 3 has been designed for this purpose. It has the single tuned filter bank, yielding a quick time response to the instantaneous signal such as burst, and has the CR smoothing network together with the LC network for the detection of envelope.

In experiment several conditions were adjusted appropriately to each material sound. The speech sound was once recorded on endless tape with clock pulse for synchronization. The analysis of the speech sound is repeated periodically. This method enables the detailed observation of

pattern and spectrum, examining the effect by the change of several conditions. The band widths of filters, type of smoothing network, the level of input sound, the frequency scale (by changing the reproduction speed) were changed. The pattern and the spectrum section were compared each other. The sampling time point of spectrum section was marked on the pattern and by observing it the sampling point was determined accurately. This procedure is necessary to observe the time change of the spectrum section in the transient signal. The conditions are as follows:

(i) Analyzing filter band width; 33 cps, 67 cps, 100 cps, 167 cps. 33 cps and 100 cps were preferably used.

(ii) Smoothing circuit after the rectification; the low pass filter for power spectrum of stationary part and the CR network for the pattern and the instantaneous spectrum of transient signal.

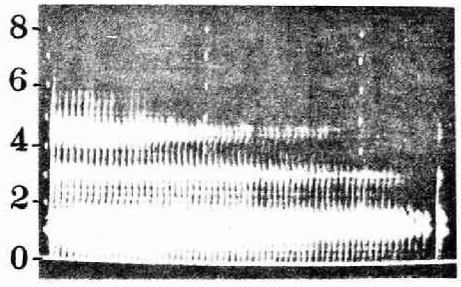
(iii) Reproduction speed of tape recorder;
normal speed for the analysis of 0—4kc range,
halved speed for the analysis of 0—8kc range,
doubled speed for the analysis of 0—2kc range.
By changing the speed, the frequency scale and the time scale are also changed.

(iv) Sampling period is 1 ms, 2 ms and 4 ms for the tape recorder speed, halved, normal and doubled, respectively.

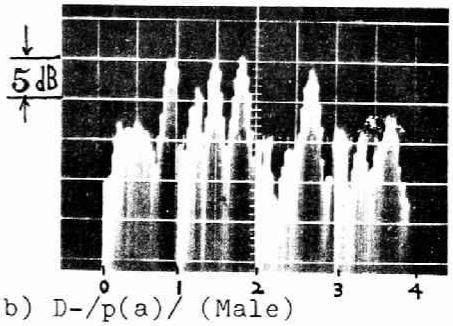
(v) The level of speech sound was adjusted to the proper level to represent the speech section of interest. When necessary, the data were taken for different levels under the same remaining conditions.

(vi) Sampling time points of spectrum sections to be photographed were determined by observing the spectrum and the pattern on which the sampling point was indicated by marker signal.

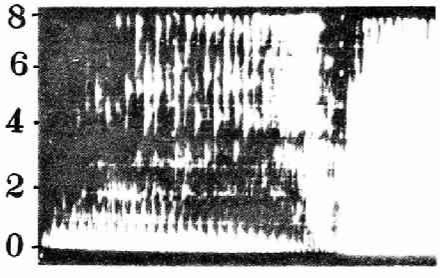
The thousands of spectra and patterns were taken. Some of these photographic data are shown in Fig.4.1. Other results are summarized in



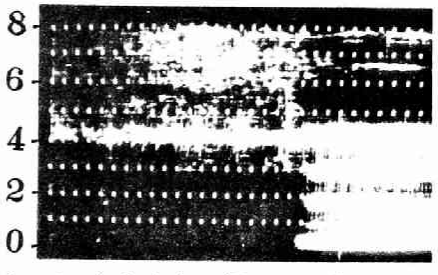
(a) S-/a/ (Female)
F_a=200 cps, CR smoothing,
Marker; 100 ms.



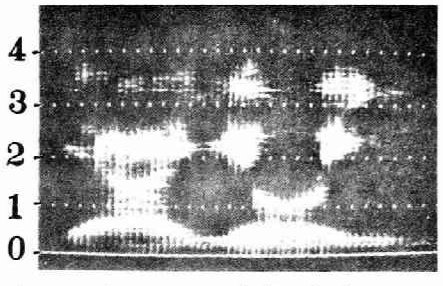
(b) D-/p(a)/ (Male)
F_a=33 cps, CR smoothing,
Sampled at 2 ms after the burst.



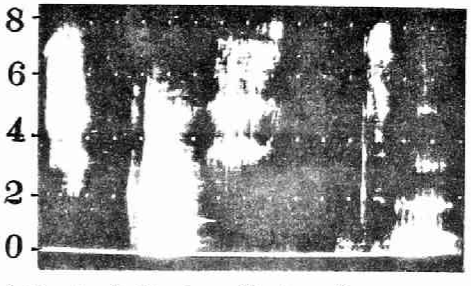
(c) S-/z(a)/ (Female)
F_a=200cps, CR smoothing



(d) G-/s(u)/ (Female)
F_a=67 cps, CR smoothing,
Marker; 10 ms.



(e) D-/ja-jo-i/ (Male)
F_a=33 cps, CR smoothing,
Marker; 20 ms.



(f) S-/tʃi-ka-ʃi-tsu/ (Female)
F_a=67 cps, CR smoothing,
Marker; 50 ms.

Fig. 4.1 Photographic data of the spectrum analysis.
(Refer to the notes of the APPENDIX I.)

APPENDIX I. (Fig.I-1 to Fig.I-8). To read these data, refer to the notes in APPENDIX I.

2. Speech Materials

Monosyllabic sounds were analyzed first to examine the parameters proper to each phoneme. The monosyllables are listed in Table 4.1 by phonetic transcriptions including the syllabic nasal and the double consonant.

In the connected speech most sounds are inter-related each other through phonetic contextual effect and are not merely the connection of monosyllabic sounds. Phonetically, phonemes are often articulated as variation according to the context (e.g., the syllabic nasal denoted by /N/ in this chapter may

Table 4.1 List of the sounds used in the analysis of monosyllables.

PHONEME GROUPE	SOUNDS	NO. OF FIG. IN WHICH DATA APPEAR
Vowel	a, i, u, e, o,	Fig. I-1, Fig. I-7
Semi-vowel	w, j,	Fig. I-2
Plosive Unvoiced Consonant	pa,pi,pu,pe,po, ta,te,to, ka,ki,ku,ke,ko,	Fig. I-3
Unvoiced Noise Consonant	sa,ʃi,su,se,so, hi,tʃi,tsu,k(i), ʃ(u), rounded and spread s and ʃ.	Fig. I-4
Voiced Consonant	ba,bi,bu,be,bo, da,dʒi,dzu,de,do, ga,gi,gu,ge,go, za,ʒi,zu,ze,zo, ra,ri,ru,re,ro,	Fig. I-5
Nasal Sound	ma,mi,mu,me,mo, na,ni,nu,ne,no, ŋa,ŋi,ŋu,ŋe,ŋo, aN,iN,uN,	Fig. I-6, Fig. I-7

be articulated as /m/, /n/, /ŋ/, /u/,etc.). Also in Japanese some additional phonemic elements must be added, such as syllabic nasal (N), long vowel, and double consonant, which are not included in the ordinary table of monosyllables and which can exist only in the connection of syllables. In the phonetic level processing the elementary sequence of recognition may consist of two or three phonemes. The longer sequences can be decomposed into these elementary sequences. The recognition is to distinguish these sequences, which must be processed as a whole, not separated to the phoneme elements. In some cases, however, the sequence is decomposed to the level of phonemes. That is, the sequences are decomposed as small unit as possible, some being decomposed to phonemes and the remainings are dealt as one unit itself. Therefore, we must first try to find the elementary recognition sequences including, of course, the phonemes.

In phonemic level, the phonemes are classified into several groups according to the manner of articulation. Each group has distinctly different properties and, within one group, the phonemes have similar structures. In the same way, in connected speech the sequences are classified into several groups. For example, the V+C+V connection and V+V connection may belong to different groups and, as for the V+V connection, /i-a/ and /i-ja/ have similar structure which must be analyzed in contrast with each other. The speech samples of the connected speech analysis must be, therefore, systematically chosen as the table of monosyllables is prepared for phonemes, although the unique selection may be impossible. The speech materials for the analysis of connected speech sound in 4.3 have been selected from such point of view. It does not contain all the items to be examined, but some of typical and important ones were picked up. For a cardinal phoneme sequence of interest, a short word including it was selected as a sample material. The words analyzed are shown in Table 4.2.

4.2 Analysis of Monosyllables

Monosyllabic sounds analyzed are shown in Table 4.1. They were classified in several groups. The results are summarized in Fig.I-1—Fig. I-7 as shown in Table 4.1. For almost sounds, patterns were taken principally under the condition B1 CR (33 cps band width of the analyzing filters and the CR smoothing network). When necessary, B1※ CR (i.e., the condition of B1 CR by reproduction speed of the tape recorder halved) was used to yield the analysis of 0—8kc range. Also, B1 LC, B2 CR, B2 LC with / without "※" were used (B2 means the condition of 100 cps band width of the analyzing filter.). The level of representation was adjusted. When the level was increased to present the weak signal, the large signal part will be over brightened.

To obtain the detailed representation, spectrum sections were taken. On the averaged spectrum of stationary part such as vowel especially on the formant measuring, many experiments and measurments have been carried out. In this chapter, therefore, the emphasis of analysis was put on the "instantaneous spectrum" and its time change for the transient signal such as the burst of consonant.

1. Vowels and Semi-vowels

The results of vowel analyses are shown in Fig. I-1 (sheet No. 1 to No. 123) of APPENDIX I for speaders D (male) and S (female).

(1). The effect of band width of analyzing filters and the smoothing network.

Patterns are shown for the condition B1 and B2, for the normal and the halved (denoted by ※) tape recorder speed and for the smoothing network of CR and LC. For male voice (speaker D), under the condition B1, harmonics are separated which is seen clearly in LC smoothing (e.g., No. 2, No. 9, etc.). This is clearer in female voice. For B1 CR of male voice, the pattern is complicated by harmonic components and vertical stripes synchronized to vocal cord excitation. When the tape recorder reproduction speed was halved

(indicated by ※), the effective band widths of analyzing filters become 67 cps and 200 cps for B1 and B2, respectively. The pitch synchronized stripes are observed in these cases (e.g.,No. 3—No.6,etc.). For high pitch frequency voice, B1 ※ shows the harmonic separation tendency, while B2 ※ shows the pitch synchronized tendency. This may suggest the extraction of pitch frequency by harmonic structure (space pattern) using a narrow band filter bank and at the same time by the pitch synchronized pattern (time pattern) using a broad band filter bank even in the female voice.

(2) Spectrum section at the initial of vowel.

The initial and the closing of the vowel sound do not always hold the regularity of the vocal cord excitation as in stationary part (e.g.,No.11, No. 31, etc.). An isolated excitation of vocal cords is often observed, though there are sounds having the regularity of excitation from onset (e.g., No. 3, No. 31, etc.). How the response of the filter bank changes with time in the onset of a vowel sound is shown by the instantaneous spectrum sections sampled at successive points, for material sound D-/o/, Nos. 48—87 and for S-/e/, Nos. 98—123., under the different conditions. In material D-/o/ at the moment of the first excitation (t=0), the section by CR smoothing has broad peaks corresponding to a formant as has been stated in chapter 2, which becomes rather sharp with time and at the next excitation it becomes again broad. In this case the harmonic structure becomes gradually dominant for the filter band width narrower than the pitch frequency. For comparison with CR, data of LC smoothing are shown in Nos. 74—87. In this condition the pattern form does not change so remarkably and the growth of harmonic structure is observed for B1. In material S-/e/, excitation in initial part of sound is irregular. In the period shown in the sections it has not reached the stationary part.

Spectrum sections of steady state are also shown with the 0-8 kc range sections for female voice. (In section representation, undesired response

may appear for the condition of the CR smoothing, because of the serial sampling in the multiplexer of analyzer as, for example, in No. 50 and No. 57.)

(3) Semi-vowels

In the Japanese monosyllable semi-vowel /j/ is followed by /a/, /o/ and /u/, and /w/ by /a/. The spectral characteristics are the gradual transitions of formants from semi-vowel to the following vowel. When input is limited to monosyllables, the detection of semi-vowel is possible.(21)(22) The problems lie in the distiction between "semi-vowel + vowel" and connected vowels, such as /i-ja/ and /ja/. In monosyllabic articulation the above distiction is not taken care of by the speaker(see 4.3, "Analysis of Connected Speech.").

The patterns and the spectrum sections are shown in Fig. I-2 for speaker D and S. As for /j/, among four formants observed in pattern the remarkable transitions are in the second and the third formants, which are also seen from a sequence of sections. At the initial part, the third formant and the fourth formant are fused into one peak whose level is stronger than that of the second formant. The second formant starts at the normal position of the front vowel, a little lower than the second formant of vowel /i/. (2 kc for speaker D, 2.5 kc for speaker S.) Before the second formant goes down, the third formant glides down to the position corresponding to the following vowel's formant. This state of formant allocation is not seen in the case of ordinary vowel. At this point the second formant begins to go down to the following vowel. The second formant transition is obvious in D-/ju/ which is articulated clearly, but in material S-/ju/ the articulation of /u/ is neutralized, that is, the movement from /j/ to /u/ is made by the lowering of the tongue keeping the mouth opening constant. Consequently the second formant is not remarkable, as is often seen in the connected speech. S-/ja/ is pronounced like /i-ja/, having the longer

duration of state /i/.

The transitions of /wa/ are observed in the first and the second formants. The initial positionsof these formants are not so low as /u/ (as for the second formant, 1 kc.). This may have been caused by the neutralized /u/, that is, the movement from /w/ to /a/ is made by the opening of the mouth cavity without appreciable change in the tongue shape.

2. Plosive Unvoiced Consonants

The analyzed results are shown in Fig. I-3. The plosive consonant (stop consonant) is composed of burst followed by the aspiration noise. Formants and antiformants are observed in the aspiration but not so distinct as in vowel. Since the plosive consonant is by no means stationary, the unique definition of power spectrum is difficult from analytical point of view. The spectrum integrated during the whole consonant duration is not suitable because the plosive part and the aspiration part must be separately processed. In smoothing the output of analyzing filter, LC smoothing may loose some time responses of the burst and the randomness of noise is also smoothed out. In this experiment, therefore, the time response to the burst signal was retained by using the CR smoothing network. In sections the instantaneous spectrum at each time point and its change with time were examined.

In the processing of the plosive consonant one problem is the distinction between the burst and the initial part of vowel. As stated in 4.2.1 the isolated pulse by an irregular excitation is observed at the onset of the vowel sound which is often confused with the burst in automatic recognition. The wave form of vowel is the damped oscillation excited behind the vocal tract, while the wave form of the burst is the transient and random impulse by the explosion of constriction of the vocal tract. The difference can be glanced from the patterns of Fig. I-3 and of Fig. I-1 for initial part of the vowel. For the vowel, even in the irregular excitation,

the section is characterized by two or three peaks corresponding to formants, but for the burst the formant is not distinct and the spectrum is often turned into the comb shaped and is split into many frequency components. This can be considered to be the effect of noise source itself or the anti-resonance by the back cavity. To clarify the structure is not easy. For comparison the responses to the rectangular wave of 2.3 ms width are shown in No. 24 and No. 25. In most cases for the filter band width B2, such feature is not so remarkable.

To see this effect clearly, the time change of spectrum sections from the burst to the aspiration was examined. Results under several conditions are shown in Nos. 44—151. It is seen in D-/po/ that, for the CR smoothing the valleys become deeper with time and comb depressions of about 500 cps spacing are shaped. For B1 CR it is clearer than for B2 and LC. But in S-/po/ this phenomenon is not seen and the spectrum section with flat envelope is obtained corresponding to the burst. In S-/te/ and M-/ke/ it is seen to some extent, but for S-/ka/ the spectrum has a formant like peak.

The spectra of the plosive consonant give different features according to the speaker, the place of articulation and the presence or absence of aspiration. As for the aspiration, /k/ which is articulated in backward position shows the formant like peaks than in /t/ and /p/. In Japanese, the plosive consonant is often pronounced with the weak aspiration, even in which case the section of burst is seen to be fairly different from the initial of vowel. Among /k/, the /k(a)/, /k(u)/ and /k(o)/ which are articulated at more backward than /k(e)/, /k(i)/, formant is dominant than the latter.

3. Stationary Noise Consonants(Unvoiced)

In Japanese, phones /s/ and /ʃ/ represent the one phoneme /s/ and the difference of both phones is not so distinctive as in English: That is, in the combination of phoneme /s/ and following vowels /a/, /i/, /u/, /e/

and /o/, the lips are rounded for /s/ followed by /u/ while the lips are unrounded for /ʃ/ followed by /i/. The situation is the same in /tʃ/ and /ts/. In the connected speech, however, the vowel following to the stationary noise consonants including affricate and aspirate is often elided. In this case the distinction of /ʃi/ and /su/, /tʃi/ and /tsu/, etc. must be made by the distinction of /ʃ/ and /s/, /tʃ/ and /ts/. In this sense /s/ and /ʃ/, /ts/ and /tʃ/ must be considered as the independent phonemes in Japanese. Phoneme /h/ is usually treated separetely as the aspirate, but for /h(i)/ the place of articulation becomes close to that of /ʃ/, yielding the similar spectrum section with /ʃ/.

The same relation as /s/, /ʃ/, and /h(i)/ is seen in the affricate /ts/, /tʃ/ and the plosive /k(i)/. In this chapter /s/, /ʃ/, /h(i)/, /ts/, /tʃ/, /k(i)/ were analyzed as one group. Further, the effect of lip rounding was examined. The patterns are shown in Fig. I-4(a), in which analysis was performed by condition B1※ CR with 0—8 kc range, and the sections in Fig. I-4(b) in which the outputs of filters of 33 cps band width are averaged by the low pass filter of 10 cps cut off.

(1) /ʃ/ and /s/.

The patterns were obtained for four speakers D, M (male) and S, G (female). To visualize weak components both large and normal level representations were taken. From patterns overall tendencies are observed which are not always expressed in sections. They show that /s/ followed by /a/, /e/ and /o/ has dominant components in fairly high region, while formants of /s(u)/ are lowered and the higher components are weak by the lip rounding effect. The formant of /ʃ(i)/ is close to that of /s(u)/ and is not so low as in English,[23] because of the unrounded lips. In some cases for /ʃ(i)/, the formant is rather high and has not so clear concentration of components as /s(u)/ which often shows the strong formant at about 4kc. The typical examples are G-/ʃ(i)/ and G-/s(u)/ (No. 53 and No. 55).

The sound G-/su/ was perceived as strongly rounded sound and it will be supposed that the above situation on formant position of /∫i/ and /su/ is caused by the rounded / unrounded of lips, that is, rounded /s/ and unrounded /∫/.

The formant frequencies of the fricative consonant are close to that of the vowel articulated by the same vocal tract configuration and anti-formants are introduced by the back cavity. By pole and zero matching on English fricative consonants, the first and the second formants are said to be damped or canceled by pole-zero pairs, showing the dominant third or higher formants.(25) That the major poles are situated in lower frequency for /∫/ and in higher frequency for /s/ is not always applied for these data in this chapter.

From the sections of Fig. I-4(b) the difference of /∫(i)/ and /s(u)/ is not clear. In English the distinction between /s/ and /∫/ in all the contexts seems to be clear, although the discrepancies of the spectra for different speakers and in different contexts are so great.(24) In Japanese /∫(u)/ is articulated with lip rounding. The spectra of M-/∫(u)/ (No. 5, No. 6), D-/∫(u)/ (No. 21, No. 22) and S-/∫(u)/ (No. 36, No. 37, No. 38) show that by lip rounding the principal peak is lowered than /s(u)/ and /∫(i)/. As for one speaker the distinction between /s/ and /∫/ by the peak position is possible to some extent, but for different speakers, especially for male and female, the common distinction is not established. (As for the detection of lower formants, refer to 2.4 of PART II.)

To examine the effect of lip rounding, articulations of /s/ and /∫/ with/without lip rounding and /s(i)/ and /∫(u)/ were made. The results are shown as patterns (Nos. 73—86) and as spectra (Nos. 56—68). The rounding of lips moves the formant to the lower frequency in both /s/ and /∫/. Rounded /s/ has higher formant than rounded /∫/ and unrounded

/s/ than unrounded /ʃ/. But the distinction of rounded /s/ and unrounded /ʃ/ is not clear from these examples.

(2) /h(i)/ and /ʃ(i)/.

Among the aspirate /h/, /h(i)/ is articulated at the advanced position near /ʃ(i)/. In Japanese both are not articulated distinctively in some cases (e.g., /hi-tʃi/ and /ʃi-tʃi/). From the results obtained, spectrum of /h(i)/ is not largely subject to change by the difference of speakers, having the major formant at about 4--5kc, while the spectrum of /ʃ(i)/ is subject to change by speakers, especially by the difference of the sex. As for one speaker, /ʃ/ and /h(i)/ are discriminable, but for many speakers not always. (e.g., D-/ʃ(i)/ and S-/h(i)/). In general, /h(i)/ is shorter in duration and smaller in level than /ʃ(i)/ but this may not be the essential difference.

(Informal perception test was tried on the output of a single tuned filter with 300--500 cps band width by changing the center frequency. /hi/ was always perceived as /hi/ but /ʃi/ was confused as /hi/ for the low center frequency about 2 kc. When this confused signal is converted to the zero-crossing signal then it was heard as /ʃi/ again. The similar phenomenon was observed between /ʃ/ and /s/. For the low center frequency of the filter, /ʃ/ was confused as /s/ though the intensity was very small and its zero-crossing version was again heard as /ʃ/.)

(3) /ts/, /tʃ/ and /k(i)/.

The initial burst of affricate can be detected in most sound, but there are some that have no remarkable burst, in which case it is apt to be confused with the fricative, though in high quality condition it is well recognized. The spectrum of /ts/ and /tʃ/ changes by speakers as in /s/ and /ʃ/. From spectrum sections it was seen that the relation between /k(i)/ and /tʃ(i)/ is similar to that of /h(i)/ and /ʃ(i)/. From patterns, however, the different features are observed. In /tʃ/ the initial burst is weak and short and after the pause of several milliseconds the

friction noise starts whose intensity is stronger than the burst. On the contrary in /k(i)/ the initial burst is stronger and the weak bursts are repeated several times, followed by the aspiration noise whose intensity is small compared with the burst. This contrast is valid for most sounds, though in M-/k(i)/ it is not clear (No. 22, No. 23 of Fig. I-4(a)). (As for the separation between /k(i)/ and /t∫ (i)/, see 3.4.1 of PART II.)

4. Analysis of Voiced Consonants

A voiced consonant is contrasted with an unvoiced consonant having the same place of articulation. Corresponding to the unvoiced plosive consonants bilabial /p/, post-dental /t/ and palatal /k/, there are voiced plosive consonants /b/, /d/ and /g/, and to the unvoiced noiselike consonants /s/, /∫/, /ts/ and /t∫/ there are voiced noiselike consonants /z/, /ʒ/, /dz/ and /dʒ/, although /dz/ and /dʒ/ are confused with /z/ and /ʒ/, respectively. In Japanese /r/ is often classified into semivowel, but here it was included in the voiced consonant, because /r/ is in general articulated as flapped in the monosyllable.

The articulation of the voiced plosive consonant is: (i) Some position of vocal tract is closed. (ii) Cavity is being closed during which the pressure of the back cavity increases and the buzz sound is radiated through wall of vocal cavity, etc.. (iii) Opening of articulation position which yields the burst, i.e., the transient random impulse with short duration. (iv) The airflow generates the fricative noise at constriction, although it is very weak and short in /b/ and /d/. (v) Vocal cord vibration is increased and formant transitions occur which tend to stationary position of the vowel. In /r/ the situation is the same as /d/, but by the flapping of the tongue at the opening of closure, burst is different from /d/ without turbulent noise. In /dz/ and /dʒ/, the fricative noise is strong and longer accompanied by a buzz sound.

In all cases, during the generation of noise, the excitation of the

vocal tract is the buzz by the vocal cord vibration. The D.C. airflow which causes the noise is modulated by the pitch frequency. At the same time the buzz source gives the vowel like sound having the formants close to the vowel of the same articulation. From these considerations the acoustic cues of voiced consonants are considered to be the buzz sound, the burst, the noise and its modulation by buzz, and the transition of formant in adjacent vowel.

The results of analysis of voiced consonants are summarized in Fig. I-5(a) and (b) ((a) for male speaker D, (b) for female speaker S). Patterns were usually taken under B1 CR or B1* CR condition. When necessary the spectrum sections were obtained for the noise part and for the burst of plosive consonants.

(1) Opposition of the voiced and the unvoiced.

The cues that characterize the voiced from the unvoiced are considered to be the presence of the buzz, rising of the pitch frequency and the first formant in the initial part of the following vowel,[26] the intensity of burst of the plosive consonant and the absence of aspiration,[27] and the modulation of noise by the buzz source in fricative consonant. These differences may be infered from the articulatory mechanism, each of which is not independent phenomenon. The basic factor may be the presence of the buzz vibration before the burst of plosive consonant, even if the presence of buzz itself is not the primary cue for the perception of the voiced sound.

As for the plosive consonant, it is seen from photographic data that the buzz is not always dominant. In some cases it is remarkable for higher formants (e.g., D-/da/ No. 11, S-/ro/ No. 59), but there are some that are not detected (D-/bo/ No. 9, S-/bu/ No. 5), which are often seen in /b/. The noise part is shorter than the corresponding unvoiced consonant and the fricative noise is weak without followed by the aspiration.

These photographs are not suitable for the detection of transition of the formant, it is, however, observed that there are some having the rising of the first formant (e.g., D-/ba/, D-/da/, S-/ra/), when it is situated at high frequency, although there are some that have not clear movement (e.g., S-/ba/). The rising of the pitch frequency toward the vowel is not clear from these data for male. For female, harmonic components show the rising of the pitch frequency in most cases. The noise part is weak and short and even it is sometimes difficult to detect in /b/, in which case the onset of the vowel has regular and quick start and the problem is the discrimination between the vowel and the voiced consonant.

The modulation of noise by buzz is observed for /g/ which has a long fricative noise duration (e.g., D-/gi/, S-/ga/, S-/gi/, etc.), even in the buzz of high frequency(female voice) as shown in spectra Nos.70--74 of S-/ga/. The burst of voiced consonant is, therefore, supposed to be synchronized with the pulse of airflow by vocal cords, which is not the case with the burst of unvoiced consonant.

The above discussion on the cues for discriminating the voiced and the unvoiced is valid for fricative consonants. For the fricative consonant, however, the sustained articulation of /z/ and /ʒ/ is possible which is clearly distinguished from /s/ and /ʃ/ without relying on the cues of the formant transition and the pitch frequency rising. The remaining factors are the presence of buzz bar and the modulation of noise by buzz.

The modulation of noise is seen in all data of /z/ and /ʒ/, especially under the condition B2*CR (200 cps effective band width of analyzing filter), whose spectrum sections sampled at successive time points are shown in Nos. 116—127 for D-/z(a)/. This may be utilized for the discrimination of voiced fricatives. As seen from these data the buzz source is excited at t=8 ms (No. 120), while the noise source is excited at t=6 ms (No. 119). There is time difference of more than 2 ms between these outputs.

In some cases buzz bar starts before the noise sound. Naturally the buzz excitation of the vocal tract generates vowel sound having the same configuration of the vocal tract as the consonant. The examples are shown in S-/za/ No. 35, S-/ʒi/ No, 38, D-/za/ No. 32, D-/ʒi/ No. 36, etc.. These formant components will serve the detection of the voiced sound together with the buzz modulated noise, even if there exists no fundamental component.

(2) Distinction of plosive consonants.

The principal cues to identify plosive consonants /b/, /d/, /g/ and /r/ are considered to be the burst and the formant transitions in the adjacent vowel. For the voiced sound the transition is more marked than in the unvoiced stop consonants and it is taken as important than the burst. The burst is, however, a main cue when the plosive is articulated, being separated from the following vowel, because the transition is not introduced. M. Halle et al. examined the perception of isolated burst and the spectral properties of it, in which /d/ and /b/, /g/ are classified by acute/grave relation and further /b/ and /g/ by the difference of levels of the major maxima of spectrum.(28)

The movement of the second formant for /b/ is flat or falling for following vowel /o/ and /u/ and rising for /a/, the starting frequency of which greatly depends on the following vowel. For the following vowels /i/ and /e/, movement is obscure. For /d/ followed by back vowels /a/, /o/ and /u/ the transition is falling, but when /d/ is followed by /e/, the difference of transitions of /b/ and /d/ is not clear. As for the third formant, for speaker D, the locus is high for /d/ and low for /g/ presenting the compact-diffuse opposition, but it is not applied to speaker S.

From these data, the distinction of /b/ and /d/ can be performed by the second formant transition when followed by a back vowel. But for the other sounds the noise burst must be considered, too. The high frequency

components of /d/ is greater than /b/ throughout the noise duration. The spectrum of noise burst of /g/ is, like /k/, subject to the influence of the following vowel. Since it is articulated in back position of the vocal tract, it has remarkable formant structure corresponding to the second or the higher formants and the duration and the level of the burst and the fricative noise are far larger than /b/ and /d/ (D-/ga/ Nos. 95—103, S-/ga/ Nos. 75—80). This compact-diffuse contrast is very useful to recognize /g/ from /b/ and /d/.

Japanese /r/ has the same place of articulation as /d/, exploded by the flapping of tongue. The formant transition is the same as /d/, but the fricative noise is not marked after the burst (D-/ra/ Nos. 138--148, S-/ru/ Nos. 93--97, S-/ro/ Nos. 98--102). When it is articulated like semi-vowel, the burst is not observed and the formant is seen during the buzz which continues to the formant of the following vowel. (S-/ra/ No, 51, S-/ro/ No. 59)

The spectral properties of the voiced fricative consonants /z/ and /ʒ/ is essentially the same as /s/ and /ʃ/ as discussed previously, except the noise modulation by vocal cord vibration and the presence of formants. What components of spectrum are due to the noise and the buzz, respectively, may be infered from spectrum sections. Spectra of D-/z(a)/ (Nos. 104—109) show the buzz components form the second and the third formants starting before the noise and that the noise components contribute in over the 4 kc region. This may be true for D-/z(u)/ (Nos. 128—132.). The result is also verified from the discrepancy of output timing of noise and buzz components as discussed above (D-/z(a)/ Nos. 116—127). The spectral properties of formants by vocal cord vibration together with the noise structure serve for the discrimination between /z/ and /ʒ/.

5. Analysis of Nasal Sounds

The vocal tract is composed of the pharynx, the nasal cavity and the

mouth cavity, which can join at the coupling point (uvular) as shown in Fig. 1.1. For periodic excitation the source is situated at glottis and the possible outputs are from the nostril and the mouth opening. For the vowel sound nasal cavity is uncoupled and the output is radiated from the mouth poening. When the nasal cavity is coupled,[29] a nasalized vowel is generated in which the nasal output is added to the oral output. The oral output then has the anti-formants, nasal formants by the nasal cavity. The oral formants for uncoupled configuration are shifted by the coupling. In the same way nasal output has the nasal and the oral formants, and the anti-formants introduced by the coupling to the oral cavity.[30] The anti-formants by the oral cavity and oral formants depend largely on the configuration of the mouth cavity and on the position of closure, which disappear in /ŋ/. Another output due to the radiation through the cavity wall is conspicuous in low frequency range. This and the fact that the wave form of vocal cords increases the lower frequency components lower the apparent first formant observed in the spectrum. Thus the spectrum of nasalized sound is complicated than the vowel by the coupling effect and further by the "addition" of two outputs (nasal and oral outputs). In general the nasal, non-vowel sound may have oral formants, nasal formants and antiformants by the oral cavity.

In Japanese there are nasal sounds; the nasal consonants /m/, /n/, /ŋ/ and the syllabic nasal N, and the vowels are often nasalized.[31] Nasal sounds have similar spectral apperance to vowel /u/, although the difference is clear with /a/, /o/, /e/ and /i/. This causes confusion in the automatic recognition of the vowel and the nasal.

The articulation mechanisms of both nasal consonant and voiced consonant are characterized by the buzz or the nasal murmur, the burst and the transition to adjacent vowel. The buzz sound generally has no nasal formants but in some cases the pronounciation like /mba/ is possible for monosyllable /ba/. The buzz of nasal consonant grows strong toward the fol-

lowing vowel, whereas the buzz of voiced plosive consonants is suppressed at the burst. The burst of nasal consonant is rather the abrupt change of transfer function than the noise components at the constriction point. From synthetic point of view it is reported that the lowering of the first formant frequency with the increased damping gives the perception of the nasality.(32)

The cue of the discrimination between nasals are said to be the transitions of the third and the second formants. The formant and the antiformant structure of the nasal murmur by the oral cavity depends on the place of articulation (i.e., /m/, /n/, /ŋ/). It also depends on the mouth cavity configuration, which and the fact that its detection from the spectrum is defficult will restrict the application of that structure to recognition. When the nasal burst is not followed by the vowel sound, remarkable formant transitions are not expected. It may suggest that the change of the spectrum properties according to the mouth opening are itself one of the important cues.

The photographic data for the sound of two speakers are shown in Fig. I-6(a) for speaker D and in Fig. I-6(b) for speaker S. In Fig. I-7 magnified spectra in the ranges 0—2kc or 0—4kc are presented, in which the effective band widths of analyzing filters are 17 cps.

(1) Nasal consonant and vowel

Nasal sounds /m/, /n/ and /ŋ/ (including /N/) are composed of the stationary nasal output which is characterized by nasal formants and, for /m/ and /n/, by oral formants and anti-formants. The spectral structures are different from vowels which are characterized by formants and also from nasalized vowels in which the oral cavity output is the major factor characterized by the nasal and the oral formants and the nasal antiformants. In general, distinction between the vowel and the nasal consonants is possible from analyzed spectrum. But for vowel /u/, the detailed analysis

must be needed. From the spectrum sections or patterns the main formants are observed at about 200--300 cps and at 2--3 kc. They do not appreciably change for /m/, /n/ and also for /ŋ/ in which coupling of the mouth cavity, consequently the oral formants and anti-formants, do not exist. For nasal consonants of speaker D the lowest is about 250 cps, the minor at 1.25 kc and the higher at 2.3 kc(which correspond to the first, the second and the third formants, respectively). The minor formant is not always observed. The distinction of zeros from formant valleys is not easy in spectrum sections. In spectra and patterns of Fig. I-6 the energy concentration of the minor formant at about 1 kc of nasal consonants is weaker than of the higher formant at about 2.3 kc. It may be seen from the spectra that the second formant of /u/ is relatively marked than the minor formant of nasal consonants and syllabic nasal, when they are compared with the third formant or higher formants, by which /u/ and nasal sounds are distinguished (e.g., D-/ŋu/, Nos. 62—66, etc.). This relation is true for speaker S(female), too.

For speaker S, the detection of formants and anti-formants is not always possible by its harmonic spacing. From Fig. I-6(b) the first formant is observed at about 250 cps, the minor formant at 1—1.2 kc when it is detectable and the higher formant at 2.5—3.3 kc ranging over the wide band. The above relation of levels of the minor and the higher formants is also seen from patterns of Fig. I-6(b)(e.g., S-/mu/, S-/nu/, S-/ŋu/, S-/uN/). Since the nasal consonant followed by a vowel is detected by the presence of the burst, the distinction between /u/ and syllabic nasal /N/ is especially important in the realization of recognition system.

(2) Burst and formant transition.

The formant positions of nasal consonants are almost constant. When the mouth cavity is opened, a dominant mouth output begins, the formants of which change according to the movement of organ. The discontinuity

of the spectral structure may cause the perception of burst. The transitions of the second and the third formants are the same as in voiced consonant,[18] in this case being observed more clearly (e.g., S-/mi/, No. 3 of Fig. I-6(b)). At the moment of mouth opening the output from oral cavity starts and the coupling to the nasal cavity is gradually diminished. Thus at the initial part of the vowel both the mouth and the nasal outputs exist, generating nasalized vowel sound. The formant situated at 2—3 kc region of the nasal consonant is seen to continue to the formant of the vowel sound. Whether it is connected to the third formant or to the second formant depends on the configuration of oral cavity at the moment of burst. From data, it is connected to the third formant of the following vowel /a/, /o/, /u/, /e/ and to the second formant of /i/. It is not always true for /ŋ/ where the oral cavity is not coupled during the consonant.

(3) Distinction between nasal consonants.

The nasal consonant is composed of the stationary nasal murmur and the burst followed by transitions, while syllabic nasal is composed of the transition from preceding vowel and the stationary nasal sound. The anti-formant structure of murmur is related to the perception of /m/ and /u/,[33] though the frequency changes by the following vowel and further no anti-formant is observed in /ŋ/. The other formants in nasal murmur do not contribute to the distinction of /m/, /n/ and /ŋ/. It may be reasonable to think that the nasal murmur chiefly serves to detect the nasality, because the phonemic distinction of the sustained nasal sound is not well perceived. Also from spectral data, the detection of anti-formant is not ensured.

The burst of the nasal consonant is the abrupt change of the vocal tract configuration. A series of spectrum sections taken near the burst are shown in Fig. I-6(a) and (b), from which instantaneous spectra of burst are observed. The overall movements of formants are, however, visualized

more clearly in patterns.

The articulation points of /m/ and /n/ are situated in an advanced position such as lips, alveolar, etc.. The speed of opening the closure is fairly fast that the burst (or abrupt change of spectrum) is easily detected in the spectrum pattern and section than in /ŋ/. In the initial part of the adjacent vowel the formants move toward the stationary positions and in some cases the nasal cavity remains coupled. The transition of formants are grouped into two, according to the following vowel. The second and the third formants of /m/ followed by back vowels /a/, /o/ and /u/ have rising transitions. For /n/, on the other hand, the second formant has falling transition while the third formant remains almost flat. Corresponding to these transition properties, it is observed from spectrum sections of the burst that the formant of /m/ followed by back vowels starts at a low frequency, so that it is not separated from the first formant (D-/ma/, D-/mu/, S-/ma/ of Fig. I-7, etc.). This property of/m/ is contrasted with /n/ in which the second formant is separated from the first formant in the spectrum sections at the burst (D-/na/, S-/na/ of Fig. I-7, etc.).

When followed by front vowels /i/ and /e/, /m/ has fast, appreciable rising in the second and the third formant, while /n/ has the flat or slight rising in these formants. These fast transitions in /m(i)/ and /m(e)/ will be smoothed when the sharp cutoff, narrow band width filter and the LC smoothing network were used in analysis.

The articulation of /ŋ/ is made at the lowered velum and the raised back tongue. The opening of the mouth cavity is started by raising velum and lowering tongue without appreciable change in the configuration of the oral cavity. By such mechanism the burst according to the opening is not conspicuous (that is, the mouth cavity opens gradually)and neither is the closure of the nasal cavity, because of the slow movement of organ (34) (In the reference the movements from resting position to complete closure

were investigated). Therefore, the initial part of vowel is much nasalized than the other nasals, /m/ and /n/. Corresponding to these mechanisms the movements of formants from /ŋ/ to vowel are not appreciable, although for /a/ the falling movement of the second formant is observed (D-/ŋa/ No.25 of Fig. I-6(a), etc.). The burst or abrupt change in spectrum is not observed, which is contrasted with /m/ and /n/ (e.g., S-/ni/ No.16 and S-/ŋi/ No. 36 of Fig. I-6(b)). Using these features /ŋ/ and /m/, /n/ followed by a vowel can be distinguished by spectra.

The characteristic properties of syllabic nasal /N/ observed from patterns and sections of Fig. I-6(a), (b) and Fig. I-7 are similar to the common properties of /m/, /n/ and /ŋ/, that is, the remarkable formant in 2--3 kc region and the weak formant at 1--1.5 kc (like the second formant of /u/). For example, the change of spectral properties from /u/ to /N/ is observed in material sounds D-/uN/ and S-/uN/. The phoneme /N/ is articulated as various phones such as /m/, /n/, /ŋ/, etc. assimilated by the articulation of the consonant, when it is followed by the consonant. The difficult problem is, however, the distinction between, for example, /u/ and /N/ in /au/ and /aN/ which is dealt in the analysis of connected speech of this chapter.

4.3 Analysis of Connected Speech

1. Introduction

The analysis of connected speech has not yet been performed systematically. One of the reasons is the complexity of phenomena by the contextual effect. Several features must be introduced other than those used in monosyllables, for examples, the duration of sound, the pause, pitch frequency, etc.. Further it must be decided how many phonemes are included in a train of speech wave. The variety of phenomena will need a large number of speech materials in the analysis. How many materials were used,

it might be impossible to cover all the varieties. The effective way is to systematically select the speech materials to be analyzed so that the intended phonetic features are clarified by analyzing them. The materials of the connected speech in this chapter were selected from this point of view as shown in Table 4.2 The perfect systematization of the connected speech is not intended here, because that will perhaps be impossible or result in increasing the complexity. The materials in this table is short form including some of the principal articles which are considered to become important at the first stage in the analysis of the connected speech.

Results of analyses are represented as patterns in Fig. I-8(a) and Fig. I-8(b) for speaker M (male) and S (female), respectively. (In the latter part of this chapter each sheet of photograph is labeled as No.8(a)35 which means No.35 of Fig. I-8(a)). The frequency range is 0—4 kc or 0—8 kc (by halving the reproduction speed of the tape recorder). The filter band width of 33 cps was used preferably, while in some cases 100 cps was also used. When necessary, a set of patterns was taken by changing the level of input speech, which is paired by link in the figure. The phonetic transcription of the sound is given under each datum in which each syllable is separated by "-". The symbols used are the same as those in monosyllables (refer to Table 4.1). New phonemic symbols N and Q were added, representing the syllabic nasal and the double consonant, respectively, whose articulations depend largely on the context. The materials were articulated a little slowly, but no other request on the speed, etc. was imposed to the speaker.

The principal subjects of analysis are also shown in Table 4.2, which are explained below.

2. Connected Vowels

Double vowel, so called here, is the connection of two vowel sounds, which is not equal to diphthong. Both vowels are equal as independent vowel in principle, but the situation is the same as diphthong in some

Table 4.2 List of materials used in Fig. I-8 for the analysis of connected speech.

Double vowels	
na-a-te	名宛て
ha-i	灰
a-u	会う
ma-e	前
ki-a-tsu	気圧
ka-o	香負
gi-iN	議員
ʒi-u	滋雨
tʃi-e	知恵
ʃi-o	塩
su-a-ʃi	素足
ru-i-ʒi	類似
su-u	吸う
su-e	末
u-o	魚
te-a-te	手当て
te-i-re	手入れ
ma-e-e	前へ
ke-u	稀有
te-o-ke	手桶
so-a-ku	粗悪
ko-i	恋
o-u	追う
ko-e	声
o-o-ki-i	大きい

Triple vowels	
mi-a-i	見合い
ko-no-a-i-da	このあいだ

Semi-vowel between vowels	
ka-ja	蚊帳

sa-jo:	作用
a-ju	鮎
i-ja	嫌
hi-ju	比喩
i-jo	伊予
tsu-ja	艶
tsu-ju	梅雨
hu-jo:	不用
he-ja	部屋

Succession of syllables of semi-vowel	
ja-ja-ko-ʃi:	ややこしい
ja-jo-i	弥生
ju-ju-ʃi-i	由々しい

Vowel, semi-vowel and consonant	
u-wa-be	うわべ
wa-ra-u	笑う
ta-ku-aN	タクアン
ja-mi	闇
i-ja-mi	嫌味
ʃi-a-ge	仕上げ
i-ʃa	医者
ʃi-ja	視野
ki-o-ku	記憶
kjo-ku	局
u-ki-jo	浮世

Elision of vowel	
ki-ʃa	汽車
ki-ko:	気候
ku-tʃi	口
ki-tsu-tsu-ki	きつつき
tʃi-ka-ʃi-tsu	地下室

so-ku-ta-tsu	速達
ku-ʃi	櫛
tʃi-tʃi-bu	父父

Consonant and double consonant	
gaQ-ko:	学校
ʃiQ-pa-i	失敗
seQ-keN	せっけん
teQ-po:	鉄砲
kiQ-te	切手
seQ-tʃi	接地
iQ-sa	一茶
reQ-ʃa	列車
i-ke	池
ki-te	来て
ha-ko	箱

Syllabic nasal and consonant	
baN-za-i	万才
saN-raN	散乱
daN-a-tsu	弾圧
huN-nu	忿怒
buN-uN	文運
a-ta-ma	頭
saN-ma	サンマ
naN-ba	ナンバ
taN-pa	短波
ʒiN-iN	人員
ʒi-niN	辞任
ʃiN-niN	新任

contexts.

The averaged duration of the double vowel is approximately 280 ms for both speakers M and S, though no instruction on speed of articulation was given for speakers.

As the parameters to describe the formant movement from one vowel to the adjacent, there are the rate of transition, the direction of movement and the continuity of the formants. As a rough approximation the formant frequencies of the vowel are classified into the several domains; as for the first formant (F_1), lower range (for /u/, /i/), middle range (for /o/, /e/) and higher range (for /a/), and as for the second formant (F_2), lower range for back vowels (/a/, /o/, /u/) and higher range for front vowels (/i/, /e/). The possible types of movements are, therefore, 9 for F_1 and 4 for F_2. As concerns the direction of movement, it is classified into the flat type, the rising type and the falling type transition. The formant does not always change continuously as seen often in the second formant, while the movement of the first formant is continuous. The second formant level of /u/ (e.g., in /a-u/, /ji-u/, etc.) are so weak that the remarkable formant energy is not observed in the spectrum, which may be caused by the neutralized articulation of /u/.

In principle movements of formants of double vowel are one directional from the starting frequency corresponding to the initial vowel to that corresponding to the final vowel. In ordinary cases, formants have the initial and the final stationary state, although durations largely depend on the context.

One of the themes in the recognition of connected vowel is the relation with the semi-vowel. The group of connected vowels from front vowel to back vowel is closely related with the semi-vowel concerning both spectrum and the perception, which may be discussed later.

3. Semi-vowel between Vowels

The semi-vowels /j/ and /w/ are characterized by transitions of formants, especially the second formant, to the following vowel's formant position.

The semi-vowel appearing in the context between vowels, such as /ka-ja/ is one of the important problems of recognition. The possible contexts are /V_1+j+V_2/ (V_2: back vowel). The movements of formants are more complicated than in the double vowel in which only the one directional glide is possible. In /V_1+j+V_2/, when V_1 is back vowel, the second formant frequency shows the low-high-low sequence and the convexing curve is observed in pattern. Since the speech materials used are pronounced a little slowly the maximum position of the second formant reaches the almost stationary position corresponding to /i/. When it is pronounced faster, it is expected that the formant can not reach the target before it comes down again. When the first vowel V_1 is front vowel /e/ or /i/, the transition of the second formant is the high-low. The change of the second formant is rapid and remarkable, ranging over the frequency about 1 kc--2.4 kc (for speaker M). On the contrary the movement of the first formant is slow and not so large. The first formant of /j/ does not reach its proper position, when it is followed and preceded by vowels with middle or higher first formant. For example, in /a-ja/ of No. 8(a)31 the minimum position is about 500 cps, which is considerably high compared with that in /a-ju/.

For the case V_1 is back vowel, the description on the movement of the second and the third formant, stated in 4.2.1 in respect of the semivowel in a monosyllable is valid for both transitions from V_1 to /j/ and from /j/ to V_2. The formant structure of the midpoint of /j/ in /a-ja/ is similar to that of /e/ rather than /i/, which is almost the case with the context having the convexing or concaving curve in both F_1 and F_2 (e.g., /a-jo/, /o-jo/, etc.). In this type of movements, the detection that there are

three phonemes in the vowel sound wave is easy, because the presence of maximum or minimum point shows the presence of the one phoneme (/j/ in this case). In this context, the discrimination between the triply connected vowel and /V_1+j+V_2/ is necessary (e.g., /sa-i-oN- i/ and /sa-jo:/,/tsu-i-o-ku/ and /hu-jo:/, etc.).

When the first vowel V_1 is /i/ and /e/, the second formant F_2 corresponding to /i-j/ stays in the higher position and the first formant in lower or the medium position. For examples in /i-ja/ and /i-jo/, the duration of stationary state like /i/ is about 120--200 ms, which is longer than in /j+V/ context.

As in the monosyllable the articulation of /u/ in /ju/ is neutralized so that the clear transition of the second and the third formant to the final state of /u/ is not seen from the pattern.

4. Succession of Syllables of Semi-vowels

The context of the connected /j+V/ syllables, that is, /j+V_1+j+V_2/, is not so frequently appear in conversation. It is, however, a very interesting context for the analysis both of the spectral relation of the semi-vowel and the connected vowel and of the mechanism of movement of articulatory organs.

The patterns are shown in No. 8(a)43--No. 8(a)45, No. 8(b)44 ——— No. 8(b)46. In context /ja-ja/ the first and the second formants are required to move within the maximum range twice (low-high-low-high or high-low-high-low). When the rate of articulation is considerably fast, the organs may be expected to move with its maximum speed. The durations of /ja-ja/ of No. 8(a)43 and No. 8(b)44 are about 300 ms. Although the speed of pronounciation is a little slow, the formants does not reach the target position peculiar to each phoneme. In speaker M (No. 8(a)43), the first state is F_1=500 cps and F_2=2.1 kc, the second state corresponding to /a/ is F_1=600 cps and F_2=1,500 cps, the third state is F_1=500 cps and F_2=2.1 kc,

then the final state is F_1=1,000 cps and F_2=1,400 cps. Thus the movement of the first formant is smoothed within the range 500-600 cps for the context /ja-j/. The second formant does not change continuously, though the movement is clear. On the contrary for speaker S, the movement of the second formant is small compared with speaker M (between 1.9--2.3 kc), while the first formant moves remarkably between 500 cps--1 kc. The context /ja-jo-i/ has high-low-high-low-high sequence of the second formant and low-high-low-high-low sequence of the first formant, in which the speech organs are requested to move a large distance. Both patterns of speaker M (No. 8(a)44) and speaker S (No. 8(b)45) have the durations of 500 ms and its movement of the second formant is clear, though the discontinuous transition occurs.

In the articulation of /ju-ju/ only the tongue position is controlled without appreciable change in lower jaw and lip opening. The second formant comes down to 1,800 cps in the intervocalic /u/ of speaker M or 2,200 cps of speaker S, which are fairly high than the final target frequency of /u/. From these data it is observed that the minimum frequency of the second formant in inter semi-vowel position is the lowest for /o/, medium for /a/ and higher for /u/.

5. Vowels, Semi-vowels and Consonants

One of the important problems concerning with the semi-vowel is the distinction between the connected vowel and the semi-vowel. As stated above, when a formant movement has the maximum point, the expectation of the presence of one phoneme is reasonable. On the other hand for contexts /ja/ and /i-ja/, since the type of formant movements is the same, the difference is in the length of the initial state corresponding to /i/ or /j/, that is, high F_2 and low F_1. The comparison of patterns No. 8(a)46, No. 8(a)47, No. 8(b)47 No. 8(b)49 and No. 8(b)50 shows that the duration of the state like /i/ is far longer in /i-ja/ than in /ja/. This is also seen in the other data.

The same situation occurs in /w/. The data having the contexts /u-wa/, /wa/ and /u-a/ are shown in No. 49--No. 52 of Fig. I-8(a) and in No. 51--No. 54 of Fig. I-8(b). The durations of initial state like vowel /u/ is longer in/u-w/ (about 140 ms) than in /w/ in which no stationary state is found. The distinctive differences between /u-a/ and /u-wa/, and between /i-a/ and /i-ja/ are not observed from the patterns (No. 8(a)47, No. 8(b)48, No. 8(b)47, No. 8(b)48).

When the semi-vowel is preceded by a consonant, /j+V/ is affected by the consonant and vice versa. For example, /ki-o/ and /ki-jo/ have the similar structure different from /kjo/, the transition of which starts at the onset of vowel part. The duration of initial state (like vowel /i/) is shortest in /kjo/, middle in /ki-o/ and longer in /ki-jo/ for both speaker M and D. A set of data of /i-ʃa/, /ʃi-ja/ and /ʃi-a-ŋe/ is also shown. In /ʃa/ the starting point of the second formant is lower than and the first formant is higher than that of ordinary /j/.

6. Elision of Vowel

Basic structure of Japanese speech sound is that it is composed of the pure vowel and the C+V syllable (C: consonant). As the additional syllables there are the syllabic nasal (denoted by N) and the double consonant (denoted by Q), each of which is thought to be one syllable. Thus C+V+N and C+V+Q connections appear. In context C_1+V+C_2, however, when C_1 and C_2 are unvoiced consonant and V is /i/ or /u/ unstressed, V is not articulated (elision of vowel). When C_2 is the plosive consonant, the silent interval is placed between the consonants. The elision of vowel also occurs in /su/ situated, for instance, at the end of a sentence. Since the vowel does not follow the consonant in the case of elision of vowel, the unvoiced consonant C_1 must be recognized without refering to the transition information toward the following vowel. /ʃ/ and /s/, /k(i)/ and /k(u)/, /tʃ(i)/ and /ts(u)/ must be identified by the structure of the consonant itself.

In context /ki-ʃa/ of No. 8(a)58 and No. 8(b)60, /k/ and /ʃ/ are continuously uttered, in which /k/ is expressed by the burst, after that the turbulent noise of /ʃ/ is articulated. The situation is the same for /ku-ʃi/, although the change of spectrum from /k(u)/ to /ʃ/ is distinct.

For the case when the *second consonant* C_2 is plosive the silent interval is placed between C_1 and C_2. The averaged duration of the silence is distributed between 50--150 ms, which is comparable with the space between the vowel and the following plosive consonant.

7. Consonants and Double Consonants

The double consonant denoted here by symbol "Q" appears in the context of preceding vowel and following unvoiced consonant. As it is articulated at the same place of articulation as the following consonant, the double consonant is often written as the succession of the consonant symbols (e.g., /gaQ-ko:/→/gakko:/). The possible consonant associated with /Q/ are /p/, /t/, /k/, /s/, /ʃ/, /tʃ/ and /ts/. The patterns are shown in No. 8(a)68--No, 8(a)78 and in No. 8(b)70--No. 8(b)77.

In double consonant context /V+Q+UVC/ (UVC: unvoiced consonants shown above), the sound of the vowel V is stopped by the closure of the mouth cavity at the place corresponding to the following UVC. The duration of the vowel is shorter than and the fading of sound is more rapid than those in normal condition. The wave form of the vowel is shown in No. 8(a)82--No. 8(a)84. The irregularity of pitch excitation and the lowering of pitch frequency, often seen in the fading section of vowel, are not observed in the context of double consonant. This is caused by the fact that the ending of sound is controlled by the closure of the vocal tract (that is, the implosion to the next consonant), not by the cease of the glottal excitation. When the UVC is plosive, the silent interval continues, the duration of which is fairly longer than in the ordinary context of vowel plus plosive consonant. The averaged duration of silence in /V+Q+plosive UVC/ is about

300 ms, while in context /V+plosive UVC/ it is about 200 ms or less.

When the UVC is fricative consonant, the noise continues about 300 ms until the next vowel starts, during which the noise structure is almost constant. For the comparison the spectra of ordinaty /V+UVC/ context are shown in No. 8(a)79--No. 8(a)81 and in No. 8(b)78--No. 8(b)80.

8. Syllabic Nasal and Consonants

Syllabic nasal denoted by "N" in this chapter appears in the context /V+N/. It is the sustained nasal sound without the burst to the vowel, even when it is followed by the vowel. The place of articulation largely depends on the following sound. That is, the phoneme N is articulated as various phones according to the context in such ways shown below:

(i) V+N (final position); closure at post-velum

(ii) V_1+N+V_2; closure at uvula

(iii) V+N+Bilabial consonant ;
(iv) V+N+Dental or Alveolar consonant;
(v) V+N+Velar consonant;
} closure at the same position as the consonant.

(In conversational speech, variation of N is more complicated.)

The natural articulation of N proper is made at various positions shown above. For V_1+N+V_2 the sudden change of timber, which is one of the cues of nasal consonants /m/, /n/ and /ŋ/, is not perceived in the transition from N to vowel V_2.

As N is articulated as different phones, its characteristics must be examined on the several contexts. The patterns are shown in No. 8(a) 85--No. 8(a)106 and in No. 8(b)81--No. 8(b)111.

As mentioned in the analysis of monosyllables, the spectral structure of /N/ is close to that of /u/, especially in final position in which /u/ is pronounced in very weak and neutralized manner. The high level of the second formant relative to that of the third formant will serve to distinguish /u/ from /N/ as seen in patterns No. 8(a)88, No. 8(b)88, etc..

The comparative study of the contexts /V_1+N+V_2/, /V_1+N+NC+V_2/ and

/V_1+NC+V_2/ (NC: nasal consonants; /m/, /n/, /ŋ/) is needed in the connected speech, for examples, /ʒiN-iN/, /ʃiN-niN/ and /ʒi-niN/. From patterns it results that the duration between V_1 and V_2 is shorter in /V_1+N+V_2/, about 100 ms, than in the context /V_1+N+NC+V_2/, in which it lasts more than 200 ms, and that the duration in /V_1+N+V_2/ has medium value. Next, by the presence of the nasal burst /V_1+NC+V_2/ and /V_1+N+NC+V_2/ are distinguished from /V_1+N+V_2/, because in the latter no burst-like change in the spectrum is anticipated as shown in No. 8(a)88, No. 8(a)91 and in No. 8(b)87, No. 8(b)106, etc..

/V_1+N+C+V_2/ context is possible for almost all the consonants C, for examples, /taN-pa/ (N is articulated in lips), /saN-raN/ (initial N in alveolar), /daN-ŋaN/ (the first N in velum), etc.. When C is the unvoiced consonant or the voiced consonant, the nasal murmur is faded or weakened between N and C, but for the nasal consonant, the steady state nasal sound continues almost constant to the vowel. The level of the formant is, however, large at the initial part of N and decreases with time.

4.4 Conclusion

Spectrum analysis was tried on the Japanese speech sounds using the device described in chapter 3. The detailed structures of results are largely affected by the characteristics of the devices. As for the formant frequencies, the observed results are similar to those obtained by the conventional spectrum analyzer using band pass analyzing filters and low pass smoothing filters. When the CR smoothing network is used, the results are rather comparable with those obtained by Sonagraph. The randomness of noise in the voiced consonant, the difference of responses to the formant and to the burst, etc. were observed from patterns and sections. The speech materials of connected speech were selected systematically so as to be able to examine the effect of context. They were classified into several groups and the relations between these groups were comparatively studied. Though

the data are not sufficient to describe the properties of all the connected sound, they give the insight in proceeding to the analysis of connected speech from the analysis of monosyllable.

Chapter 5

STATISTICS OF JAPANESE PHONEMES

5.1 Introduction

Speech recognition system will be considered at the several levels of processing; 1) acoustic analysis, 2) phonetic processing, and 3) linguistic processing. In level 1)parameter extraction is performed, the method of which will not depend on the particular language, though level 2) and level 3) need different treatments for each language. In level 2), recognition is made based on the analyzed data, considering the phonetic contextual effect (or co-articulation effect). In level 3) the phonetically recognized results (or the phoneme sequence) in level 2) are processed as a message or sentence, where linguistic informations are utilized for the improvement of score and for the final decision. The conversion from the phonemic sequence to the orthography can be performed in this level. In our daily conversation perfect understanding of speech owes the linguistic redunduncy, which complements the uncertainty in acoustic and phonetic processing. In this chapter the treatment of the connected speech in phonemic level is discussed.

It is necessary to distinguish the term "connected speech" and "conversational speech". The elementary unit of the speech sound is monosyllable, which is the input for acoustic processing. The connected speech is the sequence of phonemes where no linguistic informations are considered. This is used for the examination of the phonetic context itself without being influenced by the linguistic structure. Thus, in the processing at phonetic level, it is necessary and sufficient for the machine to recognize the phoneme sequence from the possible connected speech which is pronounced according to the phonetic system of the language and which listners can

understand. For the input of conversational speech the machine is not always required to identify the complete sequence of phonemes. In conversation the talker articulates each phoneme more roughly than he does in the utterance of random sequences, expecting that the linguistic redunduncy may be used by the listner. The uncertainty in the recognition of phonetic level may be solved by the processing of level 3).

To the recognition of connected speech, the processing must be made as a whole pattern of the speech sound, not separated to the phoneme element, by considering the overall co-articulation. There are some trials to do this by selecting as the unit of recognition the words or the spoken digits (in case of digit recognizer). Such methods may be effectively used for the limited vocabulary. For general purpose recognition it is, however, not suitable, because the number of units becomes extremely large and, also the speaker does not always pronounce the word separately.

The mutual effects of co-articulation exists even between the phonemes separately situated, but it may be simplified by taking the fact into consideration that the influence of primary importance to a certain phoneme is limited to those from the last preceding and the next following phonemes: That is, in recognizing a certain segment of a connected speech the adjacent section corresponding to three phoneme sequence must be jointly processed.

As the input sound in the processing of phonetic level the systematic speech materials must be chosen. The daily conversational speech is not appropriate to examine the structure of connected speech sound systematically and to get the knowledge for the design of machine. One way is to use some sets of samples that contain some phoneme sequences which machine must identify, by which the machine operation may be effectively examined.

For the design of machine in the phonetic level, it is necessary to consider the property of the conversational speech because the words or the sequences which frequently appear in conversation might have an importance.

As the speech sound is treated not as the word but as the three phoneme sequence in the phonetic level, the phoneme sequences which frequently appear in conversation are important.

To satisfy this requirement statistics of Japanese were obtained, in which the phonemes and several phonetic elements were selected as the elementary unit. The statistics of language have been examined in various ways.[37][39] In most of the surveys they concern with the letters or written materials, not with the phonetic aspect of spoken language, which is the knowledge needed in the processing in phonetic level. As for Japanese the statistics on Kana letter, which is the Japanese orthography, were surveyed, but very few are the statistics on the phonemic information. This information is needed in the phonetic processing of speech recognition. The processing in phonetic level is performed by the three phoneme sequence as stated above, the statistics discussed in this chapter concern with the trigram of the phonemic elements of Japanese.

5.2 Trigram of the Japanese Phoneme.[40][41]

One of the problems in the statistics is the selection of the speech materials to be surveyed. One desirable way is to use the conversation. But this may not be suitable, because it is rather rough and incomplete sentence and the vocabulary is very limited. To overcome this, in this survey materials were chosen from written texts on speech and talk. To avoid the bias of the vocabulary the journals and the novels were also used.

The next problem is the selection of symbols. In this purpose, of course, the phoneme was used. There are, however, several representations of phonemic system in Japanese. Though the basic phonemes that constitute the monosyllable were used in this survey, some additional phonemes or elements are needed in the conversational speech. The phonemes used are 23 including additional elements which are listed in Table 5.1 with the

corresponding symbol notation used in carrying out the processing. The long vowel is represented as the vowel plus long vowel symbol "L". The syllabic nasal and the double consonant were also regarded as the basic elements. In texts the space is used as the help of understanding, not of reciting the sentence. Therefore, in this investigation the space was put in appropriate points by reciting the sentense. Of course at the end of the sentence the space was always used.

The texts used consist of 50,000 phonemes in all, including the additional elements. The calculation was carried out by the digital computer. The calculation is to classify the input sequences of phonemes punched on paper tape. First, the frequency distribution N(i, j, k) of three phoneme sequence (i, j, k), i.e., trigram, was calculated, and based on the results several statistics were investigated. The original tables of trigram are presented in APPENDIX II, Table II.A. The results are the frequency of occurrence of each phoneme sequence, N (i, j, k), for the materials of 50,000 phonemes. The apperance of table is very different according to whether the first symbol is the vowel or not and the zero frequency sequences are many, suggesting the strong restriction of the phonemic structure of Japanese.

5.3 Entropy of Phoneme Sequences

The entropy of the language, H, is the average information per symbol (phoneme, here). It is, however, difficult to calculate it from the language. The approximation to H is the N-gram entropy F_n defined by Shannon.[35]

$$
\begin{aligned}
F_n &= -\sum_{ij} p(b_i, j) \log_2 p_{b_i}(j) \\
&= -\sum_{ij} p(b_i, j) \log_2 p(b_i, j) + \sum_{i} p(b_i) \log_2 p(b_i) \\
&= -\sum_{ij} p(b_i, j) \log_2 p(b_i, j) - \sum_{k=1}^{n-1} F_k
\end{aligned}
\qquad (5.1)
$$

in which: b_i is a block of N-1 letters ((N-1)-gram),

j is an arbitrary letter following b_i,

$p(b_i, j)$ is the joint probability of the N-gram(b_i, j),

$p_{b_i}(j)$ is the conditional probability of letter j after the block b_i.

The entropy H is given by the limit of F_n,

$$H=\lim_{n=\infty} F_n \tag{5.2}$$

Thus the N-gram entropy can be calculated from the table of frequency distribution of symbol(n=1), digram(n=2), trigram(n=3), etc., though the higher order statistics may not be easily calculated. From the trigram of Table II.A of APPENDIX II, the digram of Table 5.2(a), and from the distribution of symbols of Table 5.1, the entropies of the phoneme sequence of Japanese were calculated as follows:

$$\begin{aligned} F_0 &= -\log_2 \frac{1}{23}=4.524 \\ F_1 &=-\sum_i p(i)\ \log_2 p(i)=4.072 \\ F_2 &=-\sum_{ij} p(i, j)\ \log_2 p_i(j)=3.063 \\ F_3 &=-\sum_{ijk} p(i, j, k)\ \log_2 p_{ij}(k)=2.620 \end{aligned} \tag{5.3}$$

The relative entropies to F_0 are;

$$f_1=\frac{F_1}{F_o}=0.9, \qquad f_2=\frac{F_2}{F_o}=0.675, \qquad f_3=\frac{F_3}{F_o}=0.58.$$

From this, for the three phoneme sequence, the redunduncy is about 40% which is rather large than for the trigram of English letters.(36)
Also the fact that the f_2 is fairly small than f_1 suggests the strong restriction in the connection of the two phonemes in Japanese, which will be shown later.

Table 5.1

(a) Symbols used, name of phonemes and the frequency of occurrence of symbols.

SYMBOL	PGONEME*	FREQUENCY (%)	SYMBOL	PHONEME*	FREQUENCY (%)
A	/a/	13.9	N	/n/	5.2
O	/o/	11.8	R	/r/	4.1
I	/i/	9.8	M	/m/	3.1
U	/u/	6.8	D	/d/	2.7
E	/e/	6.3	G	/g/	2.0
Y	/y/	2.5	Z	/z/,/ʒ/	1.1
W	/w/	1.6	B	/b/	0.8
T	/t/	7.3	L	(1)	2.6
K	/k/	6.2	V	(2)	2.6
S	/s/	5.5	J	(3)	1.3
H	/h/	1.4	F	(4)	1.2
P	/p/	0.2			

Vowel ——— 48.6
Semi-vowel ——— 4.1

Consonant { Unvoiced — 20.6
Voiced ——— 10.7
Nasal ——— 8.3
Others ——— 7.7

Notes; * Phoneme includes some special symbols.
(1) Long vowel. Long vowel is expressed as "vowel+ long vowel symbol".
(2) Syllabic nasal.
(3) Space. Space is put when the pause may be taken in reciting the text.
(4) Double consonant.

(b) Percentage of occurrence of consonant groups.

		DENTAL AND ALVEOLAR (/t/,/d/,/n/ /r/,/s/,/z/)	VELAR (/k/,/g/,/h/)	LABIAL (/p/,/b/,/m/)	TOTAL
FRICATIVE	UVC(/s/,/h/)	5.5	1.4	-	6.9
	VC(/z/)	1.1	-	-	1.1
PLOSIVE	UVC(/t/,/k/,/p/)	7.3	6.2	0.2	13.7
	VC(/d/, /r/, /g/,/b/)	2.7 (6.8)*	2.0	0.8	5.5 (9.6)*
NASAL (/m/,/n/)		5.2	-	3.1	8.3
TOTAL		21.8 (25.9)*	9.6	4.1	35.5 (39.6)*

* Including /r/

5.4 Trigram Distribution with Rank Order

From the result of the original table of trigram, Table II.A of APPENDIX II, the rank ordered frequency of occurrence of trigram for three phoneme sequence was arranged in Table II.B of APPENDIX II. By the table, the most frequently appeared sequences "SIT" (/sit/), "YOL" (/jo:/), "DES" (/des/), etc. are the sequences corresponding to the words or the phrases that are commonly used in every conversation or sentence. The existence of such sequences in high frequency will serve the use of the redunduncy in phonetic processing of recognition. One measure of the redunduncy of the sequences is how many sequences are needed to cover some rate of all the phoneme sequences appeared in materials. From Table II.B, the rank ordered frequency curve are obtained as shown in Fig. 5.1. The decreasing of the frequency of occurrence is very fast and 600 sequences covers 80% of all the materials, and for 950 sequences, 90%. These numbers of sequences are fairly small compared with the possible number of sequences $23^3=12,167$. From these results, it may be thought that, for the recognition of phonemes in phonetic level by the context of three phoneme sequence, about 1,000 kinds of standard patterns must be prepared at least. This is not unrealizable number, but shows the possibility of realization of the contextual approach in the general purpose recognizer.

5.5 Digram and Distribution of Symbols

The digram is summarized in Table 5.2 from the trigram of Table II-A, APPENDIX II. For the value "a" of the row sum of the trigram table, the digram frequency of occurrence $N(i, j)$ was obtained as in Table 5.2 (a) which is the ordinary digram. Another possible representation is $N(i, k)=\sum_j N(i,j,k)$ as in Table 5.2(b), which is available as the column sum "b" in the trigram table. $N(i, k)$ has the correspondence to the second order transition probability.

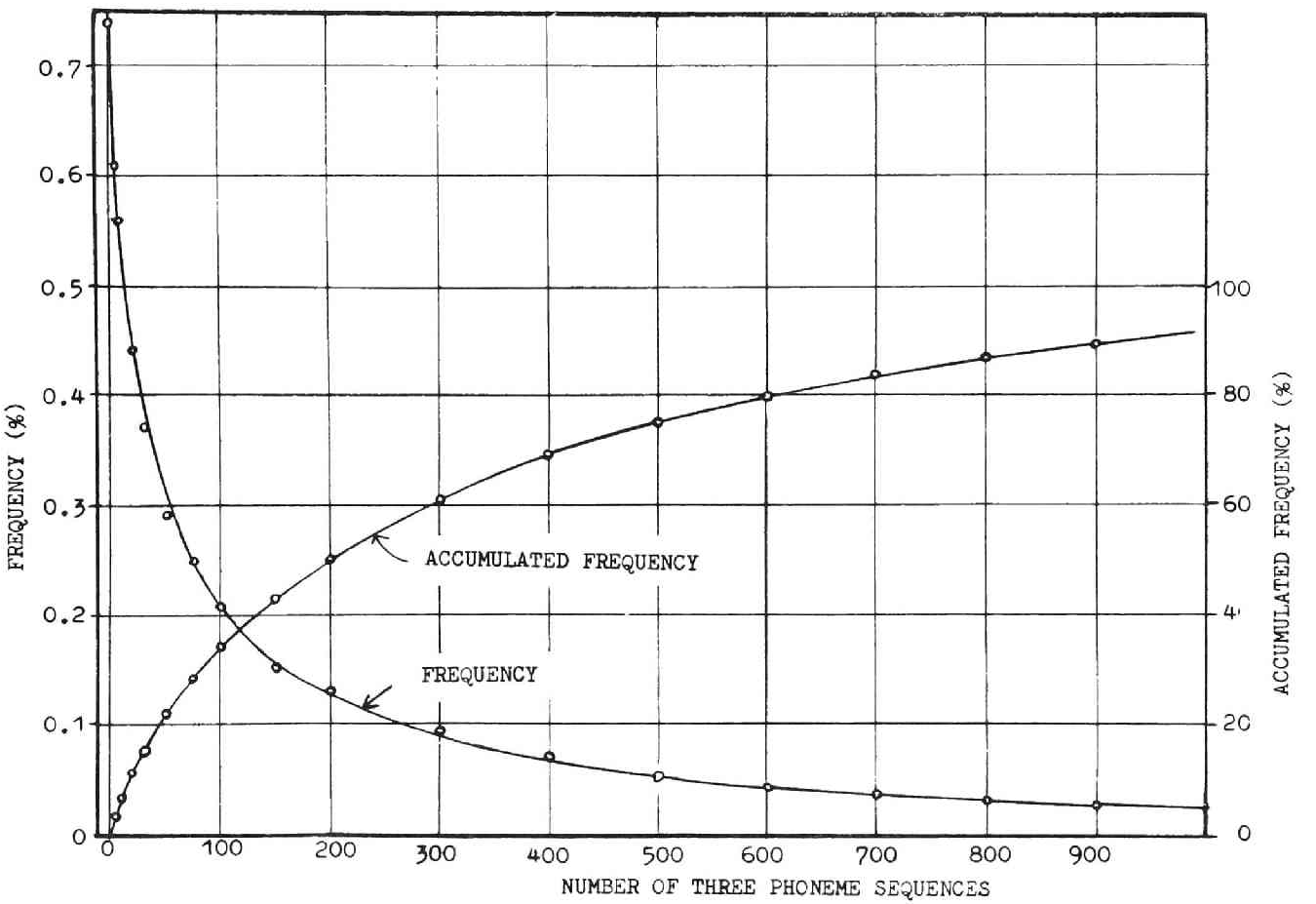


Fig. 5.1 Rank order of the most frequently occuring three phoneme sequences (trigram).

Table 5.2(a) Frequency distribution of digram of Japanese. $N(i,j,)=\sum_k N(i,j,k)$

FIRST SYMBOLS(i)	SECOND SYMBOLS (j)																						PROBABILITY OF SYMBOLS (%)	
	A	B	D	E	F	G	H	I	J	K	L	M	N	O	P	R	S	T	U	V	W	Y	Z		
A	133	43	235	116	265	193	168	696	155	844	26	279	530	126	14	813	813	623	58	434	137	93	139	13.9	A
B	159			37				28						40					113			16		0.8	B
D	372			719										275										2.7	D
E	46	51	119	13	69	66	55	453	43	252	23	178	182	70	8	192	445	346	15	313	156	48	24	6.3	E
F										63					44		63	435						1.2	F
G	711			69				67						95					54			11		2.0	G
H	195			29				162						157					101			29		1.4	H
I	60	79	242	35	140	205	110	139	25	500	21	461	533	141	12	311	321	997	10	198	160	94	119	9.8	I
J	36	8	38			11	28	23		100		46	23	18	12	6	160	88	11		28	11	15	1.3	J
K	991			290				436						691					641			63		6.2	K
L	9	19	83	13	4	87	38	28	46	179		35	225	33	5	59	181	133	3	16	34	24	59	2.6	L
M	611			149				188						512					64					3.1	M
N	721			108				657						1082					25			7		5.2	N
O	94	118	318	69	87	155	197	175	45	685	914	315	548	184	6	350	360	559	45	194	166	221	72	11.8	O
P	49			4				8						17					42					0.2	P
R	439			406				377						199					589			36		4.1	R
S	283			292				894						313					693			249		5.5	S
T	931			738				297						1189					416			78		7.3	T
U	19	55	112	28	43	147	57	114	314	386	328	164	357	41	6	280	265	360	11	104	73	63	49	6.8	U
V	5	20	213	8		142	19	10	43	101		38	204	44	21	38	119	101			35	16	84	2.6	V
W	789																							1.6	W
Y	244													632					354					2.5	Y
Z	33			37				159						23					135			174		1.1	Z

Table 5.2(b) Frequency distribution of digram of Japanese. $N(i,k)=\sum_j N(i,j,k)$

FIRST SYMBOLS (i) \ THIRD SYMBOLS (k)	A	B	D	E	F	G	H	I	J	K	L	M	N	O	P	R	S	T	U	V	W	Y	Z
A	1623	42	156	637	35	108	37	950	17	172	17	88	237	830	13	131	177	484	893	21	56	139	71
B	11	10	3	1	23	9	1	15	3	26	34	9	28	26		33	41	38	2	63	5	3	9
D	30	4	24	2	20	10	22	121	23	154	83	108	67	13	3	35	394	76	9	50	87	18	14
E	594	8	58	275	28	68	13	353	45	109	2	62	106	339	6	140	107	152	534	13	18	105	32
F	190			221				43						93					18			40	
G	61	5	43	60	20	16	36	59	5	136	25	38	66	47	5	54	98	76	22	65	29	20	21
H	21	2	34	4	22	15	6	23		60	63	14	46	10		41	16	161		90		13	32
I	1535	11	38	618	12	52	23	580		78	10	45	63	845	42	47	78	180	506	12	8	116	14
J	166	1	2	36	5		4	75		10	3	6	11	247		10	10	6	39	7	1	23	
K	23	43	156	33	131	80	32	179	26	183	107	140	348	99	2	464	229	455	32	222	62	24	42
L	316	1	16	120	6	1	8	215		16	1	7	8	340	1	5	16	29	137	2	9	58	1
M	31	6	85	37	58	32	14	92	7	75	32	46	174	43	3	98	352	179	14	59	37	10	40
N	59	50	171	32	51	92	185	163	59	344	66	127	222	77	8	128	229	170	28	117	105	59	58
O	968	24	123	536	42	81	37	673	43	194	26	99	233	1376	10	102	211	294	562	15	43	122	62
P		1	2		2	3	1			12	15	2	12	3		28	6	14		17	2		
R	30	40	117	15	18	89	58	123	17	228	38	222	193	69	4	84	169	260	24	69	94	42	43
S	57	28	42	14	23	87	37	220	301	204	61	109	208	162	8	162	146	482	63	218	43	33	16
T	67	64	133	38	37	177	58	345	73	593	98	265	313	175	8	239	186	256	19	107	140	208	50
U	696	16	71	255	3	47	35	414		155	4	52	99	651	15	34	148	157	362	21	30	85	26
V	382	1	6	219	1	2	11	149		18	1	9	7	270		3	11	14	51	3	2	100	1
W	33	6	40	2	13	13	37	38	7	136	3	37	84	13	2	72	80	118	4		17	10	24
Y	3	11	18	1	44	22	9	24	34	168	617	23	59	11	1	39	66	31	7	24	10	5	3
Z	15	19	30	5	13	6	2	41	4	45	5	19	33	117		24	30	30	57	53	8	2	3

From the tables it is observed that vowels have the affinity to almost all the phonemes though consonants to very few. As will be expected from the table of N(i, j), the distribution of N(i, k) is more uniformly distributed than N(i, j).

In the original trigram, Table II.A, APPENDIX II, the frequency of occurrence of each symbol N(i) is presented at the cell named (6). The percentages of the frequencies of symbols are listed in Table 5.1(a) and (b). When grouped according to the articulation manner, vowels occupy almost half of the total occurrences which arises from the "consonant +vowel" structure of the syllable of Japanese (about 40% in English.[37]). Also the distribution of occurrences of five vowels is rather uniform compared with that of English.[37]

Among consonants, unvoiced consonants are dominant than the voiced (plosive, fricative and flapped) and the nasal. If consonants are grouped into the transient (plosive and nasal) and the stationary (fricative), the former is more dominant than the latter which suggests the importance of the analysis of the transient signal.

As for the manner of articulation, the dental and the alveolar groups account for more than 60% of all the consonant occurrences, which is similar to the case of English.[37]

5.6 Frequency of Grouped Phoneme Sequences

To examine the structure of the phonemic sequence more clearly, the frequency of occurrence of phonemes grouped according to the articulatory manner was obtained. Table 5.3 is the grouped digram calculated from Table 5.2. From this it is seen that vowel or long vowel symbol of the first group can precede almost all the groups and the syllabic nasal can precede several groups, which is contrasted with the consonant group and semi-vowels. The pattern of this table is very remarkable, showing strong

restricitions of phonemic combinations in Japanese.

The same listing may be possible for trigram. However, it is not effective for the investigation. In Table 5.4 the frequency of occurrence to the several phoneme group sequences, which are important in the recognition of connected speech, are presented. Table (a) and (b) show that the sequences V_1+C+V_2 and C_1+V+C_2 occupy about 50% of whole frequencies, which signifies the basic structure of the Japanese speech sound. These sequences serve to simplify the recognition procedure, especially the segmentation. The difficult problem in recognition of connected speech is the processing of the vowel-like section, in which we must decide how may vowels are in the section and also whether there are semi-vowels or not; that is, the distinction of connected vowels and semi-vowels. The frequency of occurrences of such sequences is shown in Table 5.4(c), (d) and (e). The fact that they account for about 10% of the total frequency suggests the necessity of the phonetic contextual approach.

5.7 Reliability of the Results

The reliability of the statistics of language depends upon the length of materials and upon the text used. Here the effect of the length of materials is discussed.

The results obtained in this chapter are the frequency of occurrence of the trigram, the digram and the symbol. Let us n(i) be the frequency of occurrence for sequence i(i=1, 2, ...,s) for the materials of length ℓ, then the operation of calculating these statistics is to examine a set of frequencies of occurrences n(i) for each sequence in the ℓ independent trials. The probability that the frequency of occurrence for a certain sequence has the value n is expressed by the binomial distribution;

$$B(n)=\frac{\ell!}{n!(\ell-n)!}p^n q^{(\ell-n)} \qquad (5.4)$$

Table 5.3 Frequency of occurrence of digram, grouped according to the articulatory manner.

FIRST GROUP	SECOND GROUP										PROBABILITY OF OCCURENCE(%)
	V	Y	W	L	UVC	VC	NC	SN	F	J	
V	2891	519	692	1312	8389	4487	3547	1243	604	580	48.6
Y	1230										2.5
W	789										1.6
L	86	24	34		536	307	260	16	4	46	2.6
UVC	9859	419									20.6
VC	5136	237									10.7
NC	4117	7									8.3
SN	67	16	35		361	497	242			43	2.6
F					605						1.2
J	88	11	28		388	78	69				1.3

NOTES:

V; vowel. (/a/, /e/, /i/, /o/, /u/)

Y; semi-vowel. (/j/)

W; semi-vowel. (/w/)

L; long vowel.

UVC; unvoiced consonant. (/p/, /t/, /k/, /s/, /ʃ/, /h/)

VC; voiced consonant. (/b/, /d/, /g/, /r/, /z/, /ʒ/)

NC; nasal consonant. (/m/, /n/)

SN; syllabic nasal.

F; double consonant.

J; space.

Table 5.4 Frequency of occurrence of three phoneme sequences, grouped according to the articulation manner. Calculation was made on the data which has the frequency of occurrence of more than 20 in the trigram, covering 80% of the total frequency of occurrences.

V; vowel, C; consonant, SV; semi-vowel,
UVC; unvoiced consonant, VC; voiced consonant,
NC; nasal consonant.

(a) V+C+V

V+UVC+V	7854
V+VC+V	3750
V+NC+V	3408
TOTAL	15012

(b) C_1+V+C_2

		C_2			TOTAL
		UVC	VC	NC	
C_1	UVC	2850	1508	1259	5617
	VC	1604	532	434	2570
	NC	1411	500	499	2410
TOTAL		5865	2540	2192	10597

(c) $C+V_1+V_2$

C	UVC	824
	VC	392
	NC	411
TOTAL		1627

(d) V_1+V_2+C

C	UVC	657
	VC	422
	NC	357
TOTAL		1436

(e) SV

V+C+SV	297
V+SV+V	1211
V+V+SV+V	51
SV+V+V	92

in which; $q=1-p$,

p is the probability of occurrence of the sequence in the population.

To know the reliability of the experimental result n is to estimate the population mean $\ell \cdot p$ of that sequence. The distribution (5.4) has the mean value $\ell \cdot p$ and the standard deviation $\sqrt{\ell pq}$. For the value $p \leqslant \frac{1}{2}$ and $\ell p > 5$, B(n) is well approximated by the normal distribution,[38] having the same mean and standard deviation. These conditions may be satisfied in the statistics of this chapter.

Thus, for 95% confidence factor, the limits of variability of the estimated mean ℓp for the given experimental value n is calculated from the next equation:

$$p = n \pm 2\sqrt{\ell pq} \tag{5.5}$$

When $p \ll 1$, the lower limit p_m and the upper limit P_M are given by

$$P_m = (\sqrt{1+n}-1)^2, \quad P_M = (\sqrt{1+n}+1)^2, \tag{5.6}$$

and the interval of confidence limits by

$$\Delta = P_M - P_m. \tag{5.7}$$

For examples:

$n=10$, $p_m=5.4$, $p_M=18.6$ $\Delta=13.2$ $\Delta/n=1.32$

$n=100$, $p_m=82$, $p_M=122$ $\Delta=40$ $\Delta/n=0.40$

The interval becomes larger as n increases, while Δ/n decreases, giving the higher confidence. In the trigram considerable variation of the distribution will exist because the frequency of occurrence of each sequence is small. To solve this the length of materials must be increased several times. For the frequency of occurrence of symbol (Table 5.1) and of digram (Table 5.2), enough reliability may be expected.

5.8 Conclusion

In this chapter the statistical properties of the Japanese phoneme sequence were surveyed from several points. The statistics on phoneme sequence are essential for the processing of the conversational speech in phonetic level. It was known that the entropy and the frequency of occurrence of the grouped phoneme have the similar tendency to the statistics of the English phonemes. The grouped digram showed the strong restriction in the phonemic structure of Japanese. As for the trigram, the fact that about 1,000 kinds of three phoneme sequences can cover 90% of all the materials indicated the possibility of the application of phonetic contextual approach to the general purpose recognizer of conversational speech.

Chapter 6
CONCLUSION

In PART I the spectrum analysis using single tuned filter bank was discussed and in chapter 5 the statistics on the Japanese phoneme were calculated for application to the analysis in phonetic level. It was shown that the spectrum analysis by means of single tuned filter bank, together with the use of CR smoothing network in envelope detection, can represent the minute structures of speech sound that may be neglected in the conventional method using sharp cut off band pass filters and low pass smoothing networks. As the time response of a single tuned filter has fast rising up and simple structure, this analysis method is especially effective for the analysis of the transient sounds such as burst, etc..

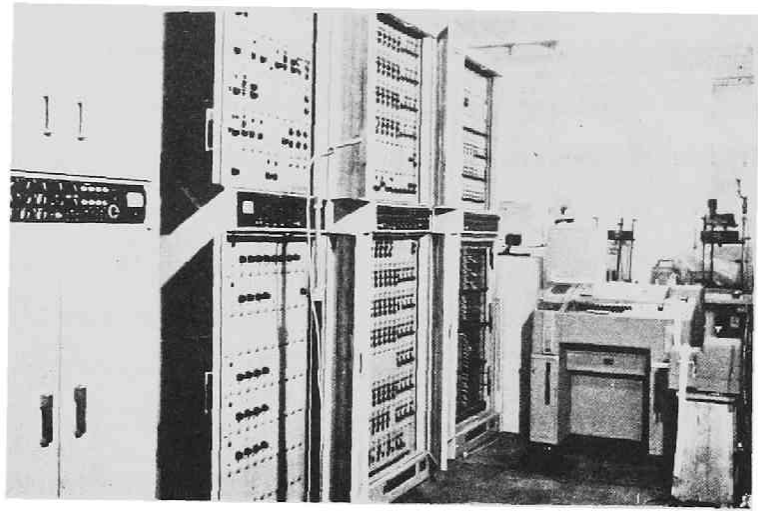
In chapter 2 the response of a single tuned filter for formant was examined and it was found that it had sufficient characteristics to the detection of formant frequency. It was calculated that in the instantaneous spectrum the formant components are expressed dominant as the time elapses, which was also observed for the speech sound (chapter 4). The analyses dealt with in PART I concern with the amplitude characteristics of the response. The analyses of the phase characteristics are stated in chapter 2, PART II.

The spectrum analyzer of chapter 3 is composed of 30 single tuned filters (effectively 240) and CR smoothing networks and the results are shown as spectrum patterns, or spectrum sections in dB scale. By observing the patterns and the sections (The sampled point of the section is marked on the pattern.), the device parameters such as the level of input sound, the frequency characteristics of the circuit, the filter band width, etc. were adjusted so as to get the desirable display.

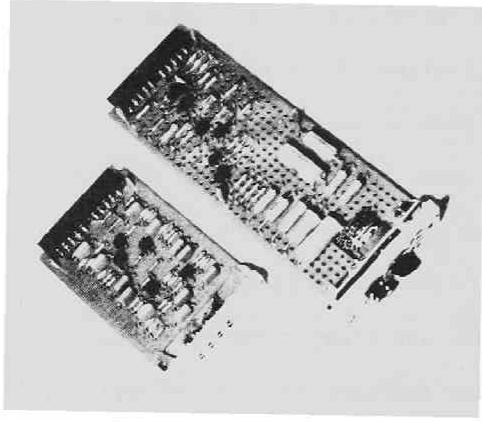
In chapter 4 analysis of the Japanese speech sound was performed using the device described in chapter 3, in which the time structure of the speech wave was examined as well as the formant structure. The randomness of noise, the periodic response pattern by vocal cord excitation, the distinction between the vocal cord excitation and the burst, the modulation of noise by pitch in the voiced consonant, etc. were also presented in the data.

The selection of speech materials is an important problem in processing connected speech. In chapter 4, the materials were selected systematically from analytical point of view as shown in Table 4.1, which were classified into several groups. The distinctive differences between these groups were observed by the analysis.

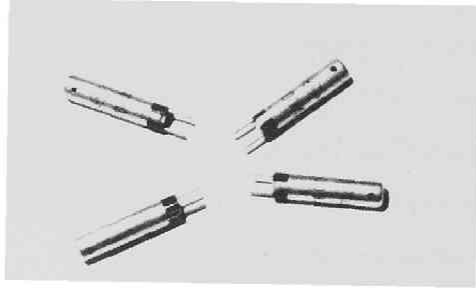
To get the knowledge on the statistical aspect of connected speech, the trigram of Japanese phoneme sequence was calculated. The phoneme was selected as the unit in considering the application to the processing of phonetic context. Among the possible three phoneme sequences of about ten thousands, about a thousand sequences could cover 90% of the whole phoneme sequences that appeared in Japanese sentences and conversations, by which the possibility of the realization of recognition system by means of three phoneme sequence (refer to chapter 4, PART II) was shown. The entropies calculated from the trigram showed the similar tendency to those of English.



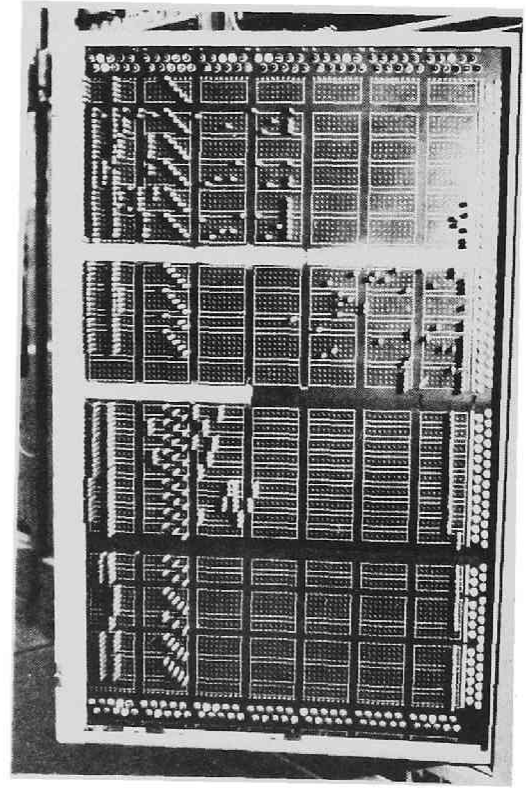
Overall view of the speech recognition system.



Package of integrating counter (upper) and shift register (lower).



Plug in diode of matrix.



Phoneme recognition matrix.

PART II

AUTOMATIC RECOGNITION OF JAPANESE SPEECH SOUNDS

Chapter 1

INTRODUCTION

One of the purposes of the analysis of the speech sound is to clarify the mechanism of the speech and to construct an automatic recognizing and the automatic synthesizing system of the speech sound. In other word it is to realize the system that can automatically perform all the functions of communication which human beings do by using speech sound. In engineering field, recognition is to detect parameters that characterize the speech sound and thereby to convert it to a sequence of codes or letters that the speech sound may transmit, and on the other hand synthesis is to generate the speech sound from a given sequence of codes or letters. Recently the problem of recognition has been systematically approached as a problem of the pattern recognition in relation to the artificial intelligence and the general principle such as statistical recognition and self-organization has been developed.

The recognition of speech sound is the problem included in these field. The general principle of pattern recognition itself, however, cannot constitute the mechanism of recognition system. The speech sound is too complicated to build the recognition system without giving any knowledge of the structure of speech. In parallel with the approach based on the general principle, the analytical study on the parameters and the linguistic struc-

ture of the speech sound are essential.

The speech sound has two kinds of informations; the linguistic information and the prosodic information. The latter relates to the naturality of speech sound that can not be translated into codes and is peculiar to the speech sound. The former is the systematic information which is represented as a sequence of codes or letters. In recognition, the linguistic information is treated.

When we regard the recognition as the conversion of speech sound to sentence or as the glasp of linguistic meaning, the processing may be classified into the levels as follows; (1) acoustic and articulatory analysis, (2) phonemic level, (3) linguistic processing. (See 5.1 of PART I.) Level (1) is to find the space and time parameters from the speech sound that characterize its phonetic colour, with which the articulatory and the perception mechanism of the human beings are related. (2) and (3) are to recognize the sequence of codes or the sentence based on the extracted parameters, according to the phonemic and linguistic structures. Among these, (3) rather concerns with the field of the linguistics, while (1) and (2) with the attribute of speech itself.

PART II of this paper concerns with the automatic recognition of the Japanese connected speech, in which the parameter extraction, the process of recognition and the processing of the connected speech by phonetic contextual relation are described. Further, the connected speech recognition system based on the above processing methods is presented, which operates in real time for the Japanese sound.

Several types of speech recognition systems have been devised. They are divided into two classes: One is the system that limits the input vocabulary,[1]-[5] usually to spoken digits, and the other does not limit the input category.[6]-[8] The system described here aims to recognize the Japanese speech sounds not limited to a certain category.[9]

One of the principal problems in recognition is the selection of the elementary unit of recognition. When the input is limited to tens of words, we can use the words as the unit of recognition by matching the input sound to these word patterns. In recognition of the connected speech, however, the variety of input category becomes extremely large as the length of the input sound increases, which makes it impossible to select words as the recognition unit. The possible way is to use phoneme as recognition unit, in which case the segmentation operation that detect each phonemic segment from the speech sound is needed. In case of processing the connected speech as a sequence of phonemes, the intereffects by phonetic contexts, which are caused by the co-articulation between adjacent phonemes, must be considered. The recognition unit must, then, be the three phoneme sequence, which method may be applicable to a general input recognition system. (See chapter 5 of PART I.)

As the parameters of speech sounds, there exist time domain parameters and frequency domain parameters. For the extraction of these parameters several analysis technics were tried, such as frequency analysis, correlation analysis, time domain analysis, zero-crossing wave analysis, etc.. None of these methods is complete for all the parameter extractions. A set of methods suitable to each parameter must be parallelly prepared.

A zero-crossing wave is a train of rectangular waves having a constant level obtained by infinite peak clipping. Compared with the 10 bits quantization of the speech sound, the amount of data to be processed in the zero-crossing wave is decreased to one tenth. The intelligibility of the zero-crossing wave was found to have high score in spite of such simplification.[10][11] The information of the zero-crossing wave is kept in the sequence of the widths of rectangular waves (i.e., zero-crossing interval).

In chapter 2 the analyses of vowel and consonant by the zero-crossing interval measurement are described, which were applied to the recognition

system.

In chapter 3 the recognition system of the Japanese sound is described. The system is composed of the recognition part and of the segmentation part. The recognition part performs analysis and recognition of speech segments operating in real time. The detection circuits proper to each parameter extraction are prepared.

Since the connected speech can contain more than one consonant and vowel, the segmentation of speech sound into sections corresponding to phonemes is required, after which each segment can be dealt by the similar way as the recognition of the isolatedly pronounced sound. In Japanese the primary importance of segmentation is in the vowel section, i.e., segmentation problem in vowel section. By controlling the decision operation of recognition system by this segmentation signal, the system can accept the connected speech.

As stated above, the phonetic context is essential phenomenon of the connected speech. The parameters which are shown when the phoneme is successively articulated are deviated from those of the separately articulated, according to the phonetic context. Therefore each phoneme can not be dealt independently, but must be jointly processed with adjacent phonemes. In chapter 4, the principle is discussed that recognizes the connected speech with phonetic contextual approach by taking the three phoneme sequence as the recognition unit. Also, the application of the principle to the connected vowel is discussed. Since this method can perform the segmentation and do the recognition of each segment at the same time, the recognition system of chapter 3 originally designed to recognize monosyllables can be extended to connected speech recognition system by adding this system.

Chapter 2

ANALYSIS OF ZERO-CROSSING INTERVALS [25]

2.1 Introduction

Speech sound is complex signal which is decomposed into frequency components having amplitudes and phase characteristics. From its amplitude response spectrum analysis is performed.

The zero-crossing wave of speech sound was found to have high intelligibility in spite of the hard amplitude limiting.[10][11] Since the amplitude of zero-crossing wave is constant, all the informations are included in the time series of rectangular waves. Zero-crossing wave represents the phase information of original signal removing its amplitude information. In the case of the sinusoidal wave phase and amplitude informations are independently separated and the zero-crossing interval holds perfect information on its frequency. However, the speech wave is the complex signal containing many components. For this reason the separate representation of phase and amplitude is not always obtained. Some conditions must be satisfied for the zero-crossing wave of the speech sound to have the information enough to recognize it.

The signal is expressed by the instantaneous frequency (or phase integrated) and amplitude. The zero-crossing wave is to observe the signal at sampled points at which the phase takes values of $n\pi+\alpha$ (n=1, 2, ...). Then, if the frequency is constant, we can estimate it from zero-crossing intervals. On the other hand, when the instantaneous frequency is too complicated to observe it at zero-crossing points, the zero-crossing intervals are of incomplete representation. To raise sampling rate, SSB signal can be utilized,[12][13][14][15][16] in which instantaneous frequency and amplitude are separately presented, and experiments of SSB clipping were applied to narrow band speech transmission systems.[16][17] The applications of zero-crossing wave to analysis or recognition of speech sound were tried in several fields.[1][18]

In every case the formants are extracted as the mean frequency of signal. To satisfy this requirement the speech signal was divided into several domains by filters so that each signal contains one formant peak.

When band pass filter is narrow enough to make spectrum analysis, the instantaneous frequency of filtered output has no information, all the informations being in its amplitude. As presented in chapter 2 of PART I, however, when the filter is of broad cut off characteristics such as single tuned filter, the phase response reflects the nature of speech signal. From the amplitude response of such filter, instantaneous spectrum is obtained and from the zero-crossings the detection of frequency of dominant components, i.e., formants can be expected, which is available for formant tracking.[19]

In this chapter the zero-crossing analysis of the direct clipped signal was performed by measuring the distribution of zero-crossing intervals, which results were applied to the speech recognition system that will be presented in the following chapter. The zero-crossing analysis of filtered signals by a bank of single tuned filters was also experimented to examine the phase aspect of the "Analysis by Single Tuned Filters" in relation to the amplitude response stated in PART I.

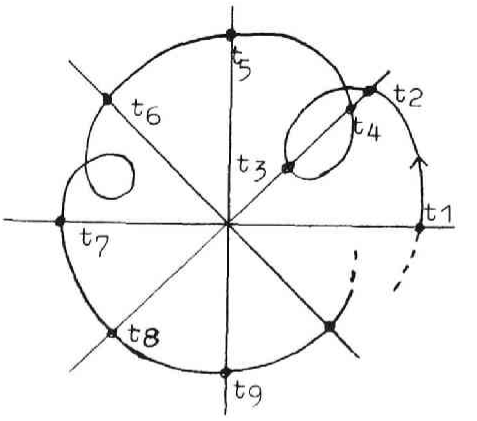
2.2 Representation of Zero-crossing Signal

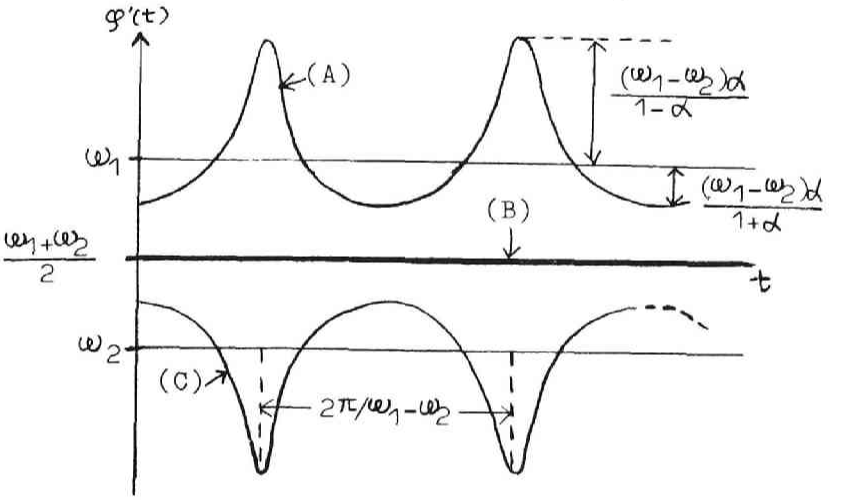
1. Representation of Signal

The zero-crossing signal is obtained by the infinite peak clipping of the speech sound. There have been adopted two conversion methods; one is the direct clipping (clipping in audio band) and another is SSB clipping of modulated signal by higher frequency carrier than the audio range. As is expected from sinusoidal wave, zero-crossing wave switches its polarity every time the vector of the analytic signal crosses some fixed axis.[13]

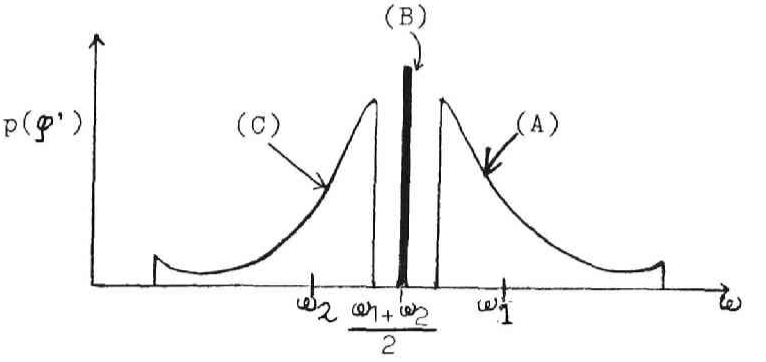
Corresponding to signal $s(t)$, which is relevant to the Hilbert transformation, the quadrature signal $\sigma(t)$ is defined as:

Fig. 2.1 Analytic signal in complex plane and its phase sampling by radial lines.





(a) Wave form



(b) Distribution

Fig.2.2 Waveform and distribution of instataneous frequency of two component tone of (2.9) , for various conditions: (A) $A_1 > A_2$, (B) $A_1 = A_2$, (C) $A_1 < A_2$.

$$s(t)=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{\sigma(t)}{\tau-t}\,d\tau\ ,\qquad \sigma(t)=-\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{s(\tau)}{\tau-t}\,d\tau \tag{2.1}$$

The analytic signal $S(t)$ of $s(t)$ is defined by expression:

$$\left.\begin{aligned} S(t) &= s(t)+j\sigma(t)=a(t)e^{j\varphi(t)} \\ s(t) &= a(t)\cos\varphi(t) \end{aligned}\right\} \tag{2.2}$$

S(t) is expressed by a rotating vector in complex plane, whose phase and amplitude are the function of time.

The amplitude of vector is expressed by;

$$a(t)=\sqrt{s^2(t)+\sigma^2(t)}\ , \tag{2.3}$$

and the phase angle by

$$\varphi(t)=\tan^{-1}\frac{\sigma(t)}{s(t)}\ . \tag{2.4}$$

The instantaneous frequency is then

$$\omega(t)=\varphi'(t)=\frac{d}{dt}\left\{\tan^{-1}\frac{\sigma(t)}{s(t)}\right\}=\frac{\sigma'(t)s(t)-\sigma(t)s'(t)}{\sqrt{s^2(t)+\sigma^2(t)}} \tag{2.5}$$

and its phase characteristics of signal $s(t)$ is

$$\cos\varphi(t)=\frac{s(t)}{\sqrt{s^2(t)+\sigma^2(t)}} \tag{2.6}$$

For a particular signal $s(t)=\cos\omega t$, $\sigma(t)=\sin\omega t$. Therefore $\sigma(t)$ of speech signal $s(t)$ may be obtained by shifting the phase by $\pi/2$ of each components consisting the speech signal.

2. Instantaneous Frequency and Zero-crossing

$\varphi'(t)$ of (2.5) can take the value of both polarities according to $\sigma's-\sigma s'\gtrless 0$. Because $a(t)\geqslant 0$, the vector makes minor loops when $\varphi'(t)<0$. Now consider to divide the plane into m sections as shown in Fig. 2.1 and represent $S(t)$ by the cross points of these axes and vector locus, then $S(t)$ is sampled by phase angle points of every $2\pi/m$. Corresponding to these phase sampling points, a set of time points

$$t_1,\ t_2,\cdots\cdots,\ t_j,\cdots\cdots,\qquad \varphi(t_j)=j\frac{2\pi}{m}\ ,\qquad j=1,\ 2,\ \cdots\cdots \tag{2.7}$$

and, using the $\{t_j\}$, a set of amplitudes

$$a(t_1),\ a(t_2),\ \cdots\cdots,\ a(t_j),\cdots\cdots \tag{2.8}$$

are sampled. As for the real signal, it is expressed as $s(t_j)=a(t_j)\cos\left(\frac{2\pi}{m}j\right)$. If the vector makes complicated locus, the sampling rate must be large (i.e., m must be large). In the particular case of $\varphi(t)$=const and $a(t)$=const, m is reduced to 1.

The zero-crossing wave corresponds to the special case of m=2, whose sampling axis is comformed with the imaginary axis at which $s(t)=0$. Therefore, if S(t) has minor loop the signal $s(t)$ is not explicitly reflected in the zero-crossing wave.

3. Example of Two Component Signal

The vowel sound has damped oscillations with different amplitudes, dampings and frequencies. For simplicity, consider two component sinusoidal wave;

$$s(t) = A_1 \sin \omega_1 t + A_2 \sin \omega_2 t \tag{2.9}$$

The amplitude, the phase and the instantaneous frequency are;

$$a(t)=A_1\sqrt{1+\alpha^2+2\alpha\cos\Delta\omega t} \tag{2.10}$$

$$\varphi(t)= \omega_1 t + \tan^{-1}\frac{A_2\sin\Delta\omega t}{A_1+A_2\cos\Delta\omega t} \tag{2.11}$$

$$\varphi'(t)= \omega_1+\alpha\Delta\omega\ \frac{\cos\Delta\omega t+\alpha}{1+2\alpha\cos\Delta\omega t+\alpha^2} \tag{2.12}$$

where $\Delta\omega = \omega_1 - \omega_2$

$\alpha = A_2/A_1$

$\varphi'(t)$ is shown in Fig. 2.2 for various conditions between amplitudes A_1 and A_2, from which it is seen that the pattern of variation depends on the relation of both amplitudes, A_2/A_1. According to $A_2/A_1 \gtrless 1$, spikes occur downward and upward, respectively, whose height becomes large as A_2/A_1 come close to 1. The spike ω_M and the trough ω_m are given by

$$\omega_M = \omega_1 + \frac{(\omega_1 - \omega_2)\alpha}{1-\alpha} \quad , \quad \omega_m = \omega_1 + \frac{(\omega_1 - \omega_2)\alpha}{1+\alpha} \qquad (2.13)$$

In particular case of $A_2/A_1 = 1$

$$\varphi'(t) = \frac{\omega_1 + \omega_2}{2} \quad .$$

One parameter of $\varphi'(t)$ is the average frequency $\overline{\varphi'}$ which is calculated as; (12)

$$\overline{\varphi'} = \frac{1}{T}\int_0^T \varphi'(t)dt = \begin{cases} \omega_1 & \text{for } A_1 > A_2 \\ \dfrac{\omega_1 + \omega_2}{2} & A_1 = A_2 \\ \omega_2 & A_1 < A_2 \end{cases} \qquad (2.14)$$

Thus $\overline{\varphi'}$ corresponds to the frequency of the larger component and the frequency of the other component is expressed by the period of the periodic pattern. The fact was utilized for the detection of dominant formant.(5)(12) Other parameter is the distribution of $\varphi'(t)$ which depends on the amplitude ratio as shown in Fig. 2.2(b). The peak point of distribution is biased toward $\frac{\omega_1 + \omega_2}{2}$ and the ranges of distributions are given in Fig. 2.2(a).

As the damped sinusoids of speech sound have different time decay constants, the ratio of A_1/A_2 changes with time in one fundamental pitch period. Accordingly waveform of $\varphi'(t)$ changes between curves (A) and (C) of Fig. 2.2(a). When ω_2 is low and A_2 is larger than and close to A_1, $\varphi'(t)$ possibly has negative value.

4. SSB Clipping and Direct Clipping

To detect precise value of instantaneous frequency of the broad band signal, the rate of sampling for the measurement of frequency must be enough high to keep the higher harmonic components of $\varphi'(t)$. For this purpose SSB modulated signal has been utilized.

The SSB signal $g(t)$ of $s(t) = a(t)\cos\varphi(t)$, modulated by the carrier of cos Wt is

$$g(t) \quad a(t)\cos(Wt + \varphi(t)) \tag{2.15}$$

In the zero-crossing wave of (2.15), the phase $\Phi(t) = Wt + \varphi(t)$ is sampled at $t_1, t_2, \cdots\cdots\cdots$ corresponding to the points at which $\Phi(t) = 2\pi, 4\pi, \cdots\cdots\cdots$. As W is chosen so high that $\varphi'(t)$ is almost constant during one interval of zero-crossings $/ t_j - t_{j-1} /$, adequate sampling is assured.[14]

To have adequate sampling points of phase angle by direct clipping, it is necessary to limit the band width of the signal to the narrower range compared with the mid frequency of the range, by which the rate of change of $\varphi'(t)$ is reduced. By making the ratio of amplitude of the dominant component to the other large, the magnitude of change of $\varphi'(t)$ can be reduced, which will serve for the good approximation of $\varphi'(t)$ by lower sampling rate.

To satisfy these conditions for speech sound, the formant is usually separated from the others by filtering. Several types of filter characteristics were tried.[1][16][17][18] The desirable condition is to contain just one formant, which is not satisfied by a fixed filter but by an electronically controlled variable filter.[17]

In the next section 2.3 fixed filters were used to detect formants, in which perfect separation of formants was not intended for the application to recognition. In section 2.4 analysis was performed on the signal passed through single tuned filter before the zero-crossing conversion.

2.3 Analysis of Zero-crossing Intervals

As has been reported by Licklider, the information of speech sound is considerably kept in the zero-crossing wave obtained by the infinite clipping of the sound wave and the articulation score is improved by differentiating the sound wave before the conversion to zero-crossing wave.[10][11] As the amplitude of zero-crossing wave is constant, the information is held in the sequence of zero-crossing points, i.e., the sequence of time intervals of rec-

tangular waves. The fact that the zero-crossing signal of the differentiated speech sound has higher articulation score than that of the original sound will show that the spectrum of the former has more spectral features of the original sound than the latter. It does not mean that these features are explicitly represented in the time intervals. As has been discussed in section 2.2, the zero-crossing signal of the two sinusoidal components (equation (2.9)) shows remarkable differences according to the ratio of A_1 and A_2 and to the difference of ω_1 and ω_2. If $A_1 < A_2$ ($f_2 < f_1$), the zero-crossing points are not sufficient to represent the higher frequency component. On the contrary when $A_1 > A_2$, the zero-crossing signal is approximated as the PPM signal whose sampling frequency is f_1, modulated by the lower frequency component f_2. Though in this case the time interval corresponding to lower component does not appear, both components f_1 and f_2 are well retained in zero-crossing wave which will give the reason to the improvement of the articulation score by differentiation.

The examples of spectra of zero-crossing signals are shown in Fig. 2.3, in which the zero-crossing signal of the two component signal $\sin 2\pi f_1 t + \alpha \sin 2\pi f_2 t$ ($f_2 < f_1$, and α in dB) is subjected to the spectrum analysis. In Fig. 2.3(a) f_1 and f_2 are considerably separated and in (b) both are closely situated. These examples show that in the case the higher component is dominant ($\alpha < 0$), the spectral components corresponding to the original sinusoidal wave are more extinguishing than in the case of $\alpha \geq 0$. This tendency is remarkable for larger separation of f_1 and f_2 (in Fig.(a)). The effect of differentiation of speech sound to the articulation score can be seen from these results.

The information of zero-crossing signal is suffered from a considerable loss by the hard amplitude clipping. The one possible approach of the zero-crossing signal is wave analysis by spectrum analysis and correlation function analysis and the other is the statistics on the time intervals of the rectan-

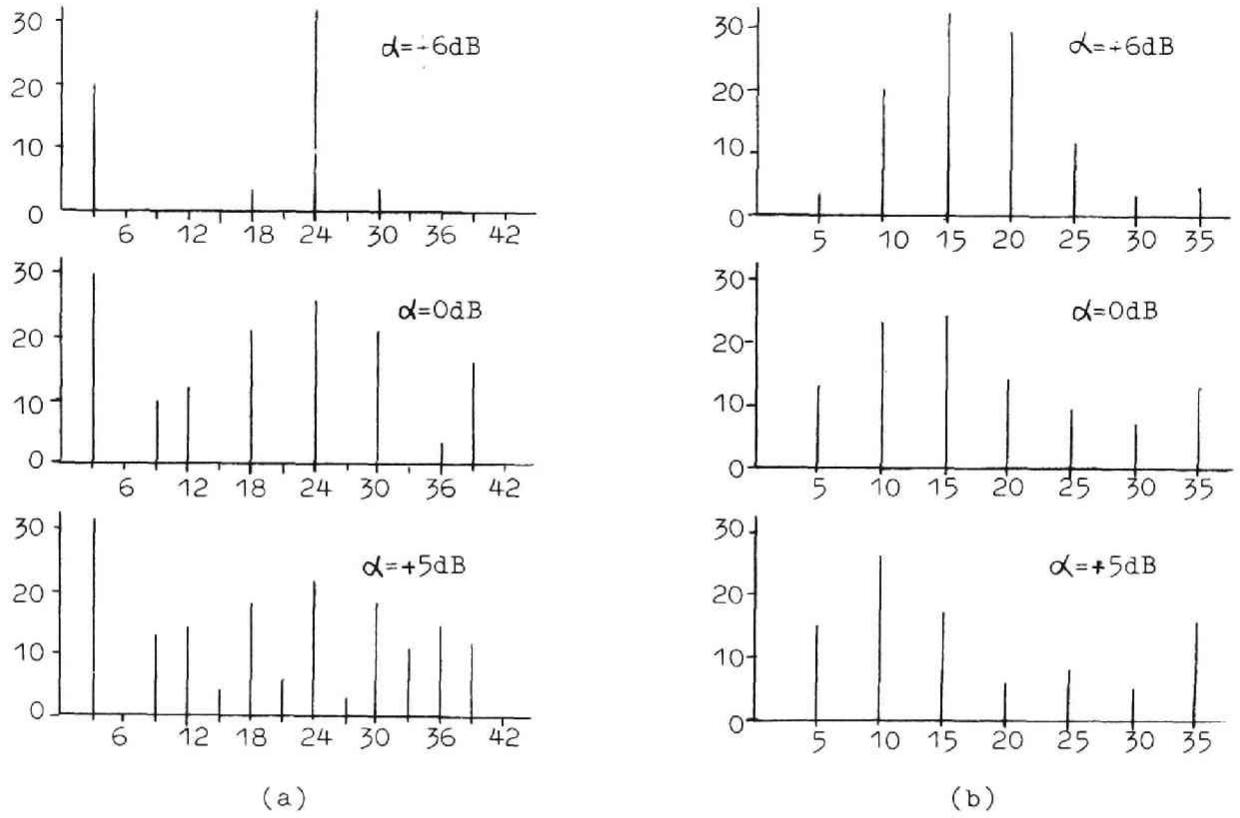


Fig. 2.3 Spectra of zero-crossing wave of two component tone, $\sin 2\pi f_1 t + \alpha \sin 2\pi f_2 t$. α is shown in dB. Abscissa is frequency(×100cps). Ordinate is amplitude of spectrum in dB. (a); f_2=300cps, f_1=2400cps (b); f_2=1000cps, f_1=1500cps.

gular waves. The merit of analyzing zero-crossing signal is in the simplicity of treatment by the dichotomization of amplitude.

As for the statistics of the zero-crossing interval, $P_n(\tau)$, the probability density of the sum of (n+1) successive zero-crossing intervals, was examined with Gaussian process having various power spectral densities.[20][21] Corresponding to $P_0(\tau)$, probability distribution of zero-crossing interval $W(\tau)$ of the speech sound in regard of long term statistics was experimentally obtained by Davenport.[21] Another possible expression is $P(\tau_1, \tau_2, \cdots\cdots)$, the joint probability distribution of the successive intervals τ_1, τ_2, $\cdots$. The higher the order of joint probability distribution, the more precise expression of the zero-crossing signal is possible, although the complexity of treatment of the signal will increase rapidly with the order of the distribution.

The zero-crossing analysis presented in this paper is the probability distribution, which will be called zero-crossing distribution hereafter. The distribution obtained by Davenport concerns with the long term property in which speech sound was taken as an ergodic process. What is required in the analysis of speech sound is, however, to examine the structure of the elementary part of speech sound and must be short time or running expression. In this paper the zero-crossing distribution was obtained by measuring and classifying the zero-crossing intervals into several channels and by averaging it in a short time enough to separate each phonemic segment consisting the speech sound.

1. Definition of Zero-crossing Distribution

Let's consider a series of rectangular waves obtained by the infinite clipping of the speech sound and let's $\tau(t)$ a width of rectangular waves that exists in some time point t . The probability that $\tau(t)$ becomes

$$\tau_i < \tau(t) < \tau_i + \Delta\tau_i$$

is given by

$$W(\tau_i)\Delta\tau_i = \text{Prob.}\{\tau_i < \tau(t) < \tau_i + \Delta\tau_i\} \qquad (2.16)$$

This is not identical with the distribution of intervals $P_0(\tau)$, but it concerns with the distribution of the width of the interval that exists in a time point t. $W(\tau_i)$, i= 1, 2, ...n, will give a zero-crossing distribution classified to n channels.

As in the analysis we treat the distribution summed in short duration of speech sound, the time averaged distribution is not equal to the ensemble averaged distribution defined by equation (2.16). But hereafter we dare treat the time averaged distribution for short time duration.

By Davenport,[22] the time averaged distribution of equation (2.16) was given by the next equation.

$$W(\tau_i) = W_i = \lim_{T\to\infty}\lim_{\Delta\tau_i\to 0} \frac{N_i \tau_i}{T\Delta\tau_i} \qquad (2.17.a)$$

in which N is the number of occurrences of rectangular waves having the interval $\tau_i \sim \tau_i+\Delta\tau_i$, which appeared during the period of interest T. ($\Delta\tau_i$ is assumed to be small.) $W(\tau_i)\Delta\tau_i$ expresses the rate of summed period of rectangular widths, whose intervals are in $\tau_i \sim \tau_i+\Delta\tau_i$, to the total duration T. Corresponding to this, the frequency distribution of zero-crossing intervals is defined as follows:

$$W(\tau_i) = \lim_{T\to\infty}\lim_{\Delta\tau_i\to 0} \frac{N_i}{T\Delta\tau_i} \qquad (2.17.b)$$

In experiment the limit $T\to\infty$ and $\Delta\tau_i\to 0$ of equations (2.17.a) and (2.17.b) are not realized. Therefore we omit these operations. Also, the expressions of (2.17.a) and (2.17.b) are substantially equal. By such assumptions, the zero-crossing distribution is expressed as

$$W(\tau_i) = W_i = \frac{N_i \tau_i}{T\Delta\tau_i} \quad , \quad i= 1,2,\cdots\cdots,i,\cdots,n. \qquad (2.18)$$

in which $\Delta\tau_i$ is the width assigned for the i-th channel, τ_i is its center frequency and n is the total number of channels.

2. Method of Zero-crossing Analysis

The zero-crossing distribution defined by equation (2.18) can be obtained by several methods. The principal procedures required are; the conversion of speech to zero-crossing signal, the measurement of the zero-crossing intervals, its classification to channels and the summation of results. Davenport measured the intervals of zero-crossing interval using amplitude modulation and level selector. Sakai and Inoue[23] measured the interval by digital method and the results were integrated by counters. They also tried the interval classifier by cascaded monostable multivibrators.

The method used in this paper is based on the digital method and the representation of distribution is made on a set of counters and by a set of voltages.

The blockdiagram of zero-crossing analysis circuit is shown in Fig. 2.4. Speech sound is first passed through filters, the characteristics of which will largely affect the results of zero-crossing analysis. There are several methods to obtain the zero-crossing version of the signal. One is the cascade connection of difference amplifiers. By this method the signal is clipped step by step in several stages, not clipped at one time. If the input wave is distorted or unsymmetric, after the incomplete clipping in intermediate stage, its zero level will be shifted from that of the original signal. Therefore, after the clipping of several stages the zero-crossing points are not identical to those of the original signal. To avoid this effect the signal was amplified to several hundreds volts by linear amplifier and then it was clipped at one stroke by a diode clipper. Finally clipped signal was shaped to rectangular wave. The dynamic range of clipping is about 60 dB.

Before the measurement of interval, full-half selection and polarity selection are performed. As the interval of rectangular wave, we can use τ^+,

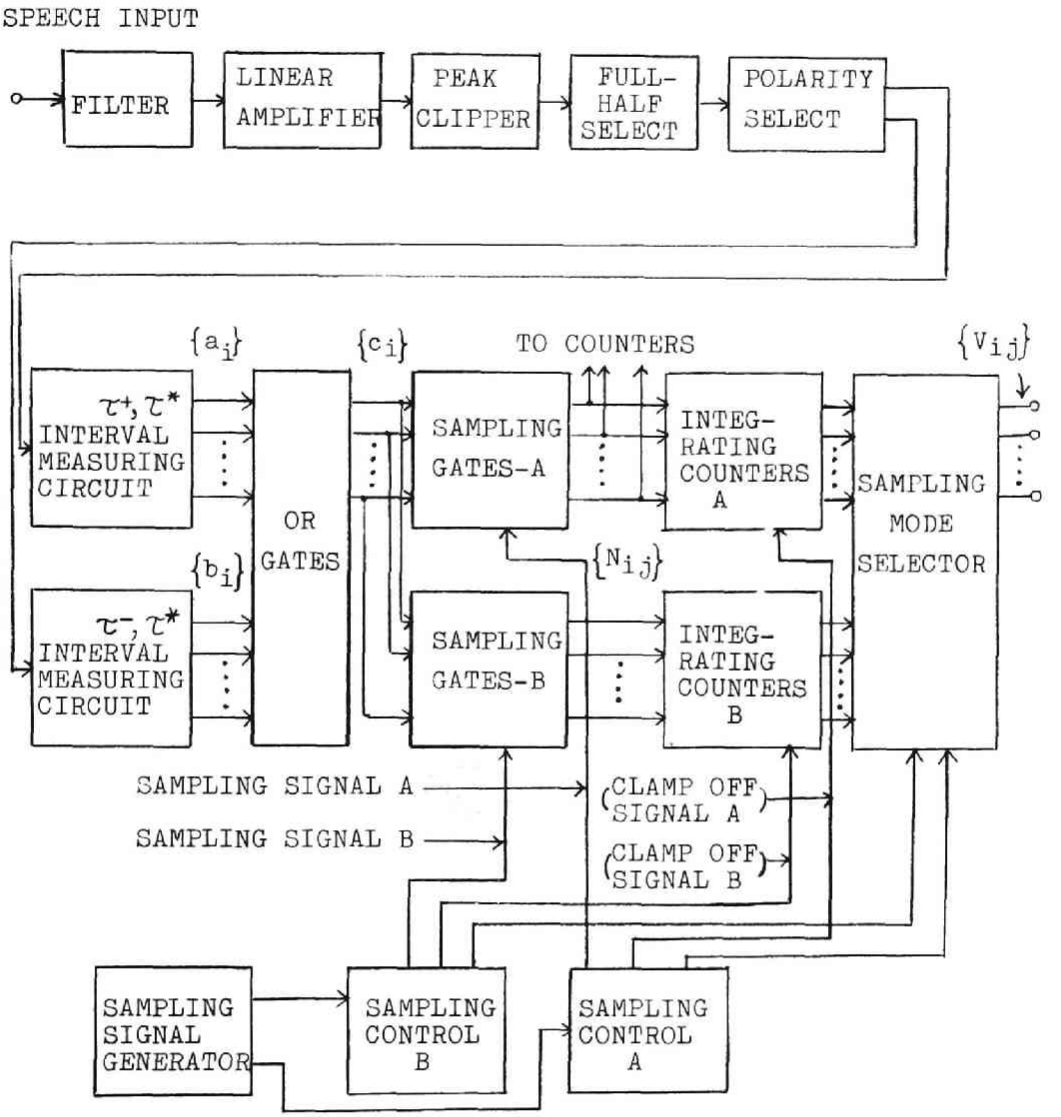


Fig. 2.4 Blockdiagram of zero-crossing analysis circuit, operating under the successive sampling.

τ^- and τ^*. τ^+ and τ^- are the half cycle width of the rectangular wave during which they stay in positive and negative level, respectively. τ^* is the full cycle of rectangular wave, i.e., the interval between the adjacent positive (or negative) going zero-crossing points. In the case of distorted sinusoidal wave, τ^* will give correct estimation of its frequency rather than τ^+ or τ^-. But, for example, for the signal that consists of two components having comparable amplitudes, τ^* will miss the detection of interval of lower frequency components, being troubled by a short interval introduced by higher components. Three kinds of polarity selections are possible; τ^+, τ^- and the mixture of τ^+ and τ^-.

By these reasons, there are prepared a pair of measuring circuits of zero-crossing intervals. One is for τ^+ and another is for τ^-. The constitution of both circuits is the same. The block diagram is shown in Fig. 2.5 and its operation is in Fig. 2.6. In the illustration of Fig. 2.5, τ^+ is measured. At the positive going of zero-crossing signal, the pulsed Hartley starts its oscillation, whose frequency f_c is chosen enough high than the signal frequency, and continues it until the next negative going zero-crossing comes. The output is shaped to clock pulses. The width τ_k^+ are measured by counting the number of pulses in a train in such way as $\Delta d_k \leqslant \tau_k^+ < \Delta d_k + 1$ (see Fig. 2.6). The counter outputs are connected to decoding matrix consisting of programmable diode logics, in which counted number d_k is classified into one of the several channels i (i= 1, 2, ... , n), according to the logics set by diodes. The decoding matrix is excited at the end of each rectangular wave to be measured. One pulse appears in the i-th channel such as $p_i \leqslant d_k \leqslant p_{i+1}-1$, where p_i, p_{i+1} are the lower bounds of i-th and (i+1)-th channel, respectively. Just after the matrix is excited, counters are reset preparing for the next measurement. The outputs from two measuring circuits, $\{a_1, a_2, \cdots, a_n\}$ for τ^+ measurement and $\{b_1, b_2, \cdots, b_n\}$ for τ^- measurement, are gathered by OR circuit, generating $\{c_1, c_2, \cdots, c_n\}$.

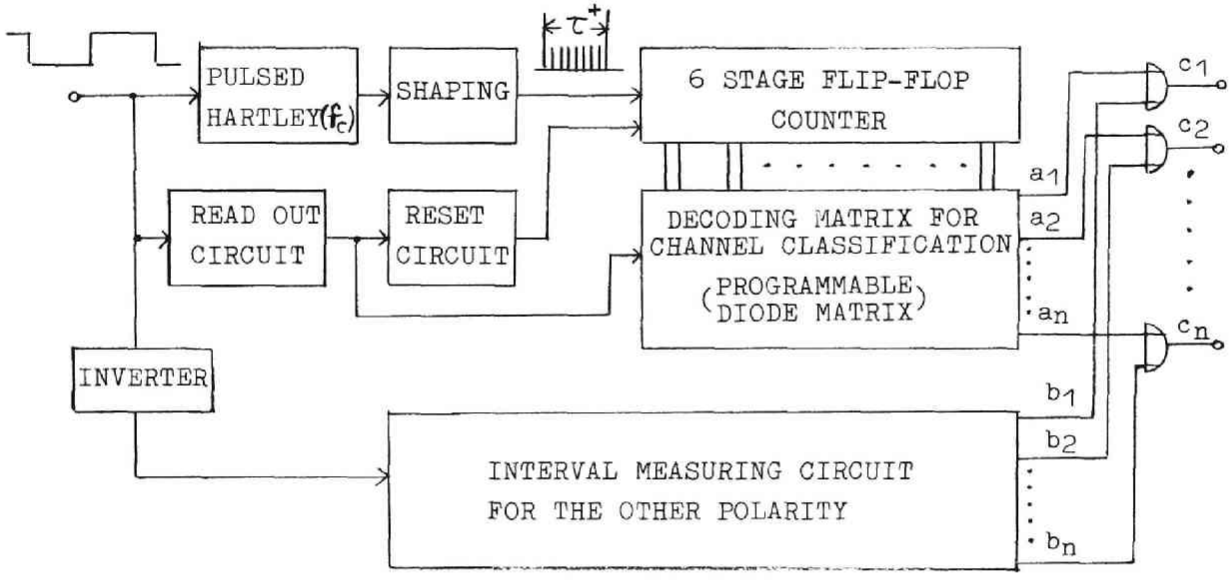


Fig. 2.5 Blockdiagram of the zero-crossing interval measuring circuit.

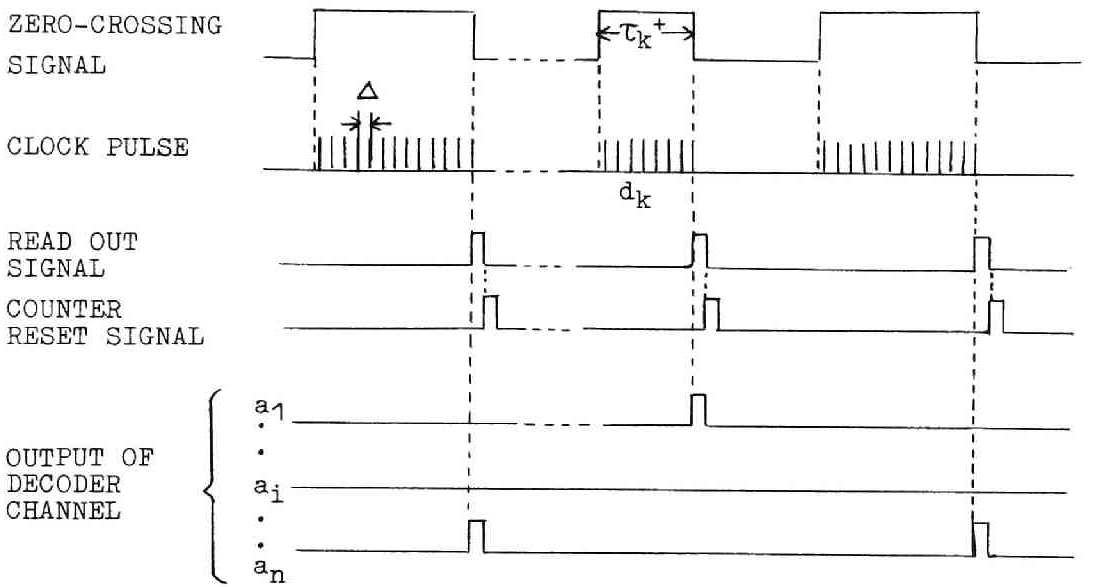


Fig. 2.6 Operation of zero-crossing interval measuring circuit for the measurement of width τ_k^+ with positive polarity. $\Delta d_k \leqq \tau_k < \Delta(d_k+1)$ for k=1,2,·····, where Δ is the period of clock pulse, typically 50μs. a_i=1 at the moment of read out signal just after the τ_k, for $p_i \leqq d_k \leqq p_{i+1}-1$.

The measured intervals $\{c_1, c_2, \cdots, c_n\}$ are sampled by sampling gates. In general, the sampling signals T_1, T_2, ..., T_j, ... are repeatedly applied to obtain a series of short time zero-crossing distributions which bear the time variation information of speech. For the j-th sampling, we can obtain the distribution of zero-crossing numbers $\{N_{ij}\}$ $(i=1,2,\cdots,n)$, corresponding to the $\{W_j(\tau_i)\}$ of equation (2.18). If the sampling is made only once, $\{N_{ij}\}$ can be read on a counter. For the successive samplings counter is inappropriate. Another representation is to convert N_{ij} to the voltage V_{ij} proportional to it and then represents $\{W_j(\tau_i)\}$ as a pattern. For this purpose integrating counters were used.

The circuit to obtain the voltage V_{ij} from N_{ij} for the successive sampling T_1, T_2, ..., T_j, ... is shown in Fig. 2.7, and its operation is in Fig. 2.8. To make the successive sampling sampling intervals are divided in two modes, A and B, which are repeated in such way as A, B, A, B, ... For each mode a sampling gate, an integrating counter and an output gate jointly operate under the control of sampling signal, clamp off signal and mode select signal. The operations of both circuits are the same except the timing of the control is complementary in A and B, as shown in Fig. 2.8.

Sampling signal generator sends a periodic sampling signal of e.g., 20 ms under the control of input speech sound. The signal is divided by flip-flop. The control signal A and B are led from each side of the switch. Sampling signal opens the gate and pass the pulse fed from zero-crossing interval measuring circuit. The monostable multivibrator generates a pulse of a certain width. At the same time clamp off signal is applied and release the clamp of the integrating circuit of R and C. The rate of charging up per one input pulse is varied by selecting capacitor and by adjusting the pulse width of the monostable multivibrator. After the gate is closed, the voltage accross capacitor is kept constant, which represents the V_{ij} for the i-th channel and in the j-th sampling period. The V_{ij} for mode A is sent to OR gate through AND gate

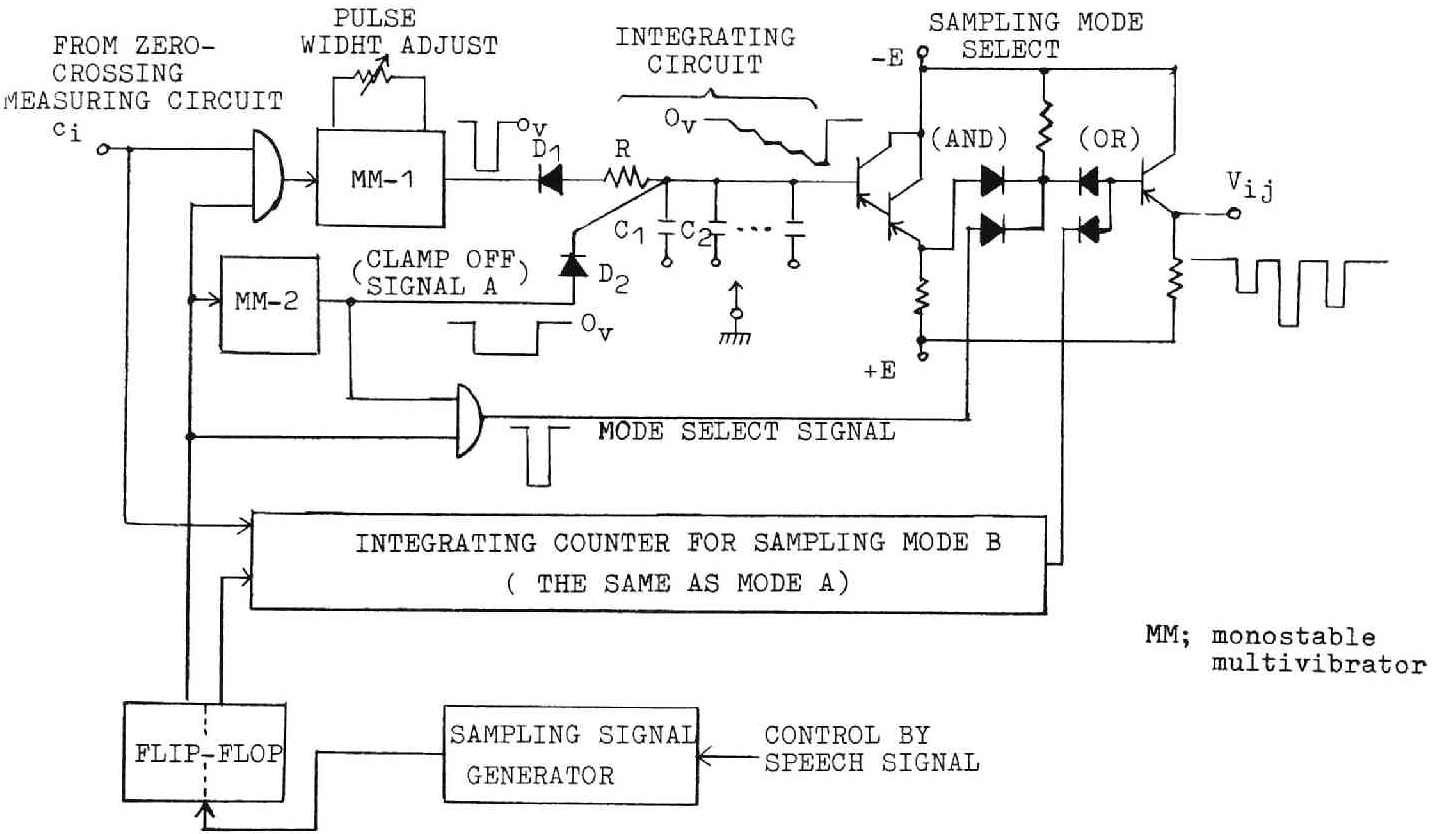


Fig. 2.7 Integrating counter and sampling mode selector for the conversion of the pulse number of the classified zero-crossing distribution into the proportional voltage. The sampling is repeated periodically(e.g., T=20ms) under the control of speech signal to be analyzed. The figure is shown for mode A.

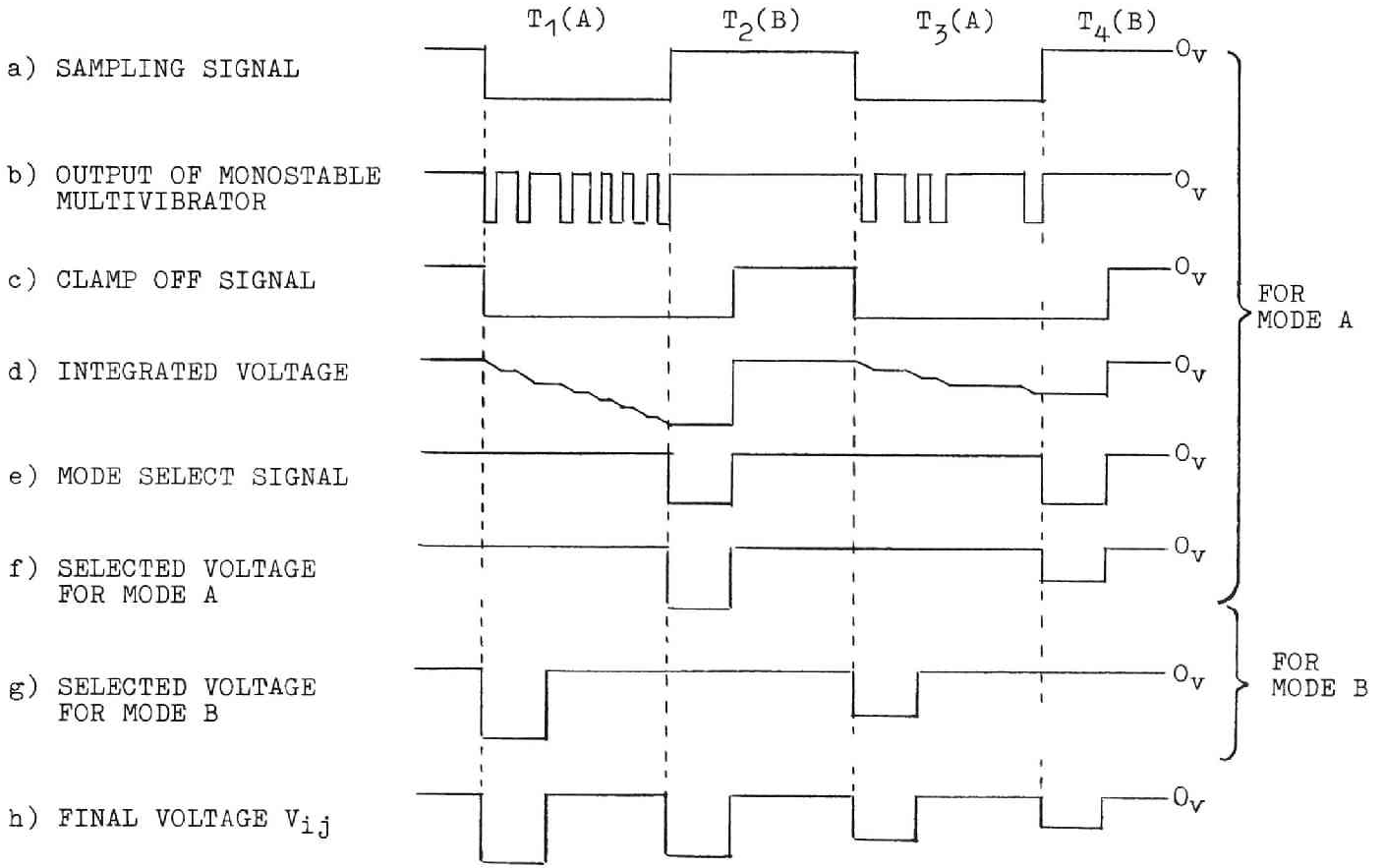


Fig. 2.8 Operation of the integrating counter and the sampling mode selector. The successive sampling intervals T_1, T_2,$\cdots$ are applied alternativelly to circuit A and B. Circuit B operates complementally with A.

controlled by mode select signal. After the OR gate combined the output voltages of mode A and of mode B, the final output is, then, the V_{ij}.

The voltages $\{V_{ij}\}$ thus obtained will correspond to zero-crossing distribution $\{W_j(\tau_i)\}$ or $\{W_{ij}\}$, i=1, 2, ..., n and j=1, 2, ... , j, ... In obtaining $W_j(\tau_i)$ from $\{N_{ij}\}$, appropriate weighting by τ_i and $\Delta\tau_i$ is needed as shown in equation (2.18). The arbitrary weighting can be realized in the integrating counters by selecting the conversion ratio from N_i to V_i as stated above. In analysis, exact normalization of equation (2.18) was not always intended, but it was adjusted to the appropriate value for each purpose.

2.4 Zero-crossing Analysis of Speech Sound by Single Tuned Filter

The information of zero-crossing wave is in the time points at which the complex signal crosses the imaginary axis. When the signal contains one formant the amplitude and the phase can be independently expressed. The phase of signal composed of more than two formants, however, suffers from mixed influences from formant frequencies and amplitudes. In previous section the signal was simplified to contain one formant or the cluster of formants by passing it through a band pass filter, on which zero-crossing analysis was made. Since it is impossible to separate the formant perfectly by fixed band pass filter, the results obtained is the approximation of formant frequency, although the electronically controlled variable filter can avoid this overupping formants to fall in one filter band.[18] In the zero-crossing analysis of the signal passed through band pass filter, it is necessary that the signal contains at least one formant to avoid unwanted output by weak inter-formant components.

Another selection of the type of filter is possible, which have a broad cutoff characteristics. The simplest is a single tuned filter. The output of this filter contains formant components, even when formants exist at the detuned position from filter pass band, though it suffers attenuation proportional to the distance between the frequencies of the formant and the filter. If

there is no formant just tuned to the filter, the zero-crossing wave of smaller components will be produced. On the contrary, when one formant is tuned to the filter, other weak components will be neglected in the zero-crossing wave. The defects of wide band, sharp cutoff filter are removed in single tuned filter. The use of single tuned filter was tried in the formant tracking using mean frequency.[19]

In this section the zero-crossing distributions of the formant signals passed through a single tuned filter are examined and next the formant extraction method based on that principle is proposed. The experiment was performed on the vowel and the consonant.

1. Phase Response of Single Tuned Filter to Formant and its Zero-crossing Wave

The phase response of the single tuned filter (with center frequency F_a and the band width B_a) to the formant shaped signal (with formant frequency F_f and the band width B_f) may depend on the parameters F_a, B_a, F_f and B_f. The output $e_o(t)$ of the single tuned filter shown in Fig. 2.9 is given by equation (2.5a) and (2.5b) of chapter 2, PART I, the amplitude, the phase and the instantaneous frequency being given by (2.10)-(2.12). Since $e_o(t)$ consists of one component when $F_a=F_f$, we treat the case $F_a \neq F_f$ in this section. The relation of amplitudes of both oscillations is decided by B_f and B_a which changes with time. The ratio of amplitudes at time t after the impulse was applied is given by

$$\alpha(t) = \frac{A_f}{A_a} = \alpha_0 \exp[-\pi(B_f - B_a)t] \qquad (2.19)$$

in which: A_f is the amplitude of the formant component, $\omega_f = 2\pi F_f$

A_a is the amplitude of the filter component, $\omega_a = 2\pi F_a$

$$\alpha_0 = -\frac{\sqrt{1+(B_f/\omega_f)^2}}{\sqrt{1+(B_a/\omega_a)^2}}$$

suffix a and f relate to filter and formant, respectively.

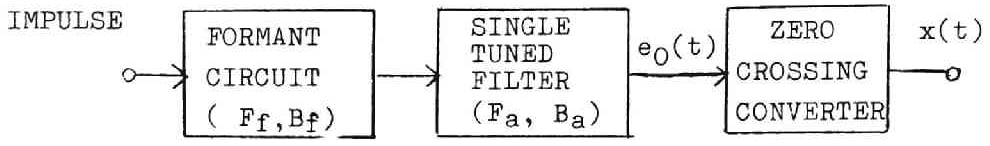


Fig. 2.9 Block diagram of zero-crossing analysis with single tuned filter.

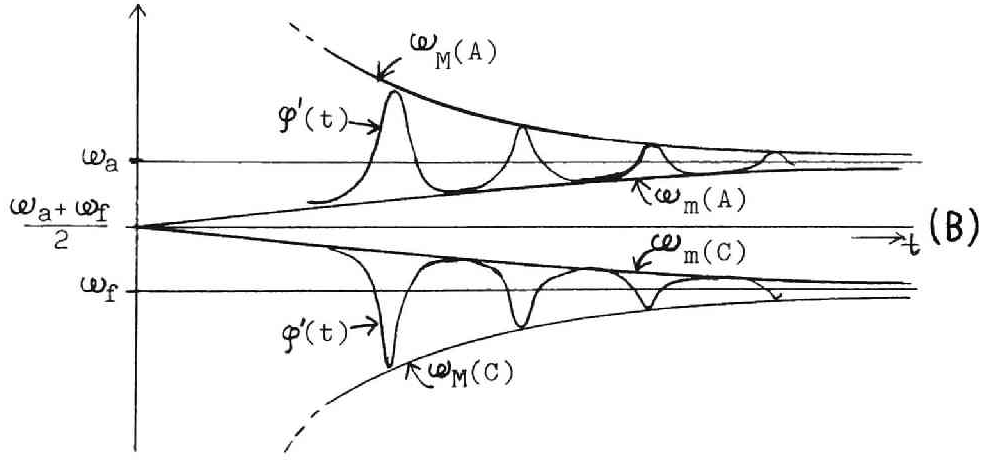


Fig. 2.10 Inatantaneous frequency and its maxima and minima of the signal $e_0(t)$ passed through single tuned filter. Condition: (A) $B_a < B_f$, (B) $B_a = B_f$, (C) $B_a > B_f$.

Since the band widths of formant and filter are fairly smaller than the center frequencies, we put $(B_f/\omega_f)^2 \ll 1$, $(B_a/\omega_a)^2 \ll 1$. Then $\alpha(t)$ is approximated as

$$\alpha(t) \doteqdot \exp(-\pi \Delta B t), \quad \Delta B = B_f - B_a \tag{2.20}$$

By using this $\alpha(t)$, peaks and troughs of instantaneous frequency given by (2.13) will change with time as shown in Fig. 2.10. It is seen from this figure that, when $B_f > B_a$ (i.e., broader formant than analyzing filter), instantaneous frequency $\varphi'(t)$ tends to ω_a and when $B_f < B_a$ to ω_f, while for $B_a = B_f$ it stays in $(\omega_a + \omega_f)/2$.

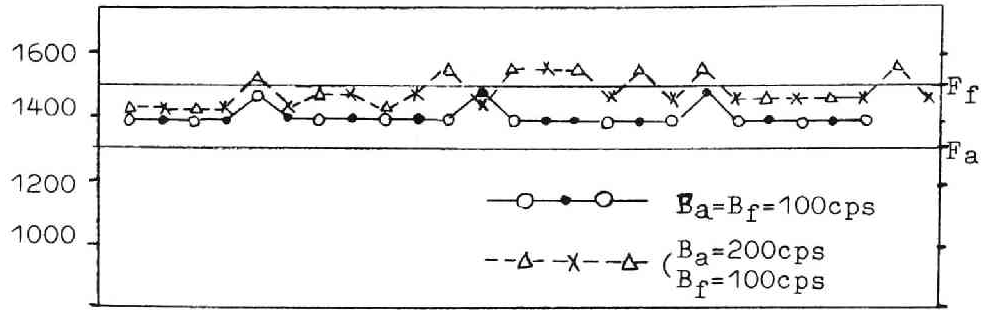
The series of zero-crossing intervals can be estimated from these $\varphi'(t)$ patterns which oscillate with the period of $2\pi/|\omega_f - \omega_a|$. The calculation of the zero-crossing intervals was carried out. One of the results is shown in Fig. 2.11 as a time series of zero-crossing intervals. The distribution of zero-crossing interval shows similar pattern as the distribution of $\varphi'(t)$.

Zero-crossing distribution can be obtained by integrating the time series of intervals. As is seen from Fig. 2.11, the distribution will depend on the duration of integration for the case of $B_a \neq B_f$. When the circuit of Fig. 2.9 is excited by periodic impulses integration more than one period will be required.

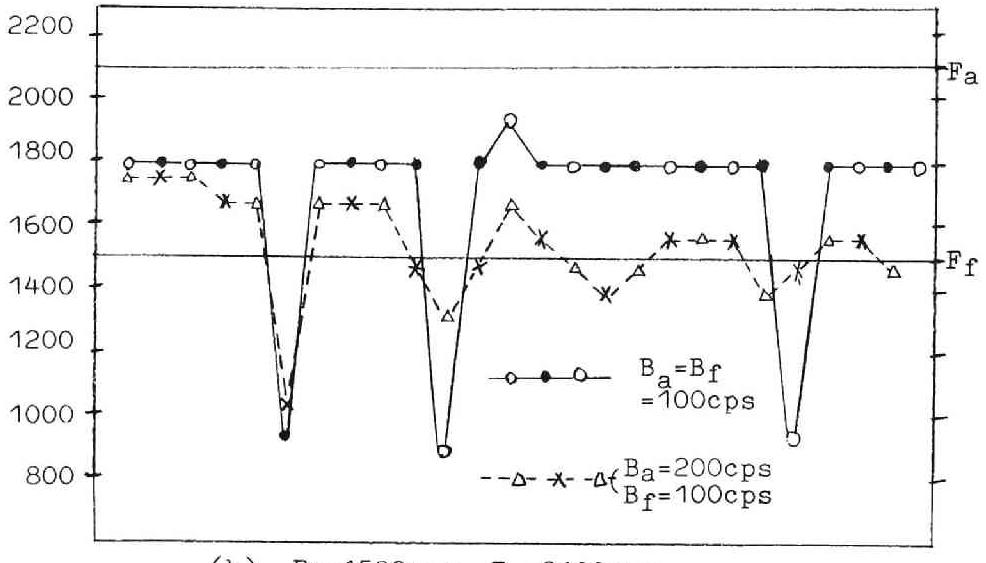
Distribution of zero-crossing intervals $W(f)$ is shown in Fig. 2.12 to see the effect by the difference of center frequencies of formant and filter and by the difference of band widths of formant and filter. The integration was made for one fundamental period.

2. Formant Extraction by Zero-crossing Analysis Using Single Tuned Filter

The mean frequency $\overline{\varphi'(t)}$ (or abbreviated as $\bar{\varphi}$ hereafter) of $e_0(t)$ of Fig. 2.9 coincides with the frequency having narrower band width, F_a or F_f. Now suppose the band width of filter B_a is shosen wider than that of formant B_f, then $\bar{\varphi} = F_f$ is satisfied. Therefore, by measuring $\bar{\varphi}$ of the signal passed through single tuned filter having broader band width than for-

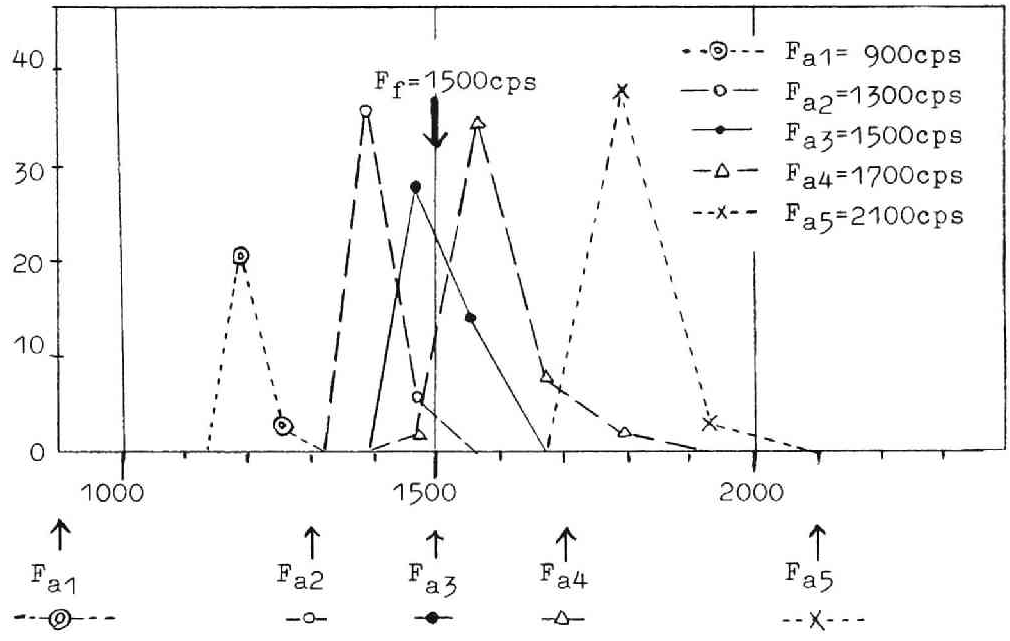


(a) F_f=1500cps, F_a=1300cps

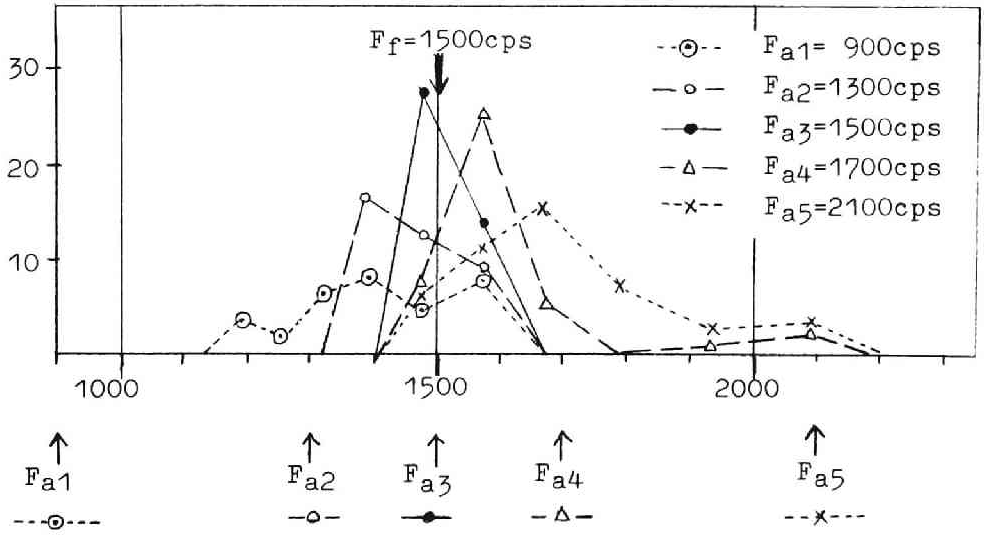


(b) F_f=1500cps, F_a=2100cps

Fig 2.11 The sequence of zero-crossing intervals of the formant signal with frequency F_f and band width B_f passed through a single tuned filter with center frequency F_a and band width B_a. Formant circuit was driven by single impulse. Ordinate is frequency in cps which is the inverse of the measured zero-crossing interval. o and •, Δ and X indicate the polarity of the rectangular waves.



(a) $B_a=B_f=100$cps



(b) B_a=200cps B_f=100cps

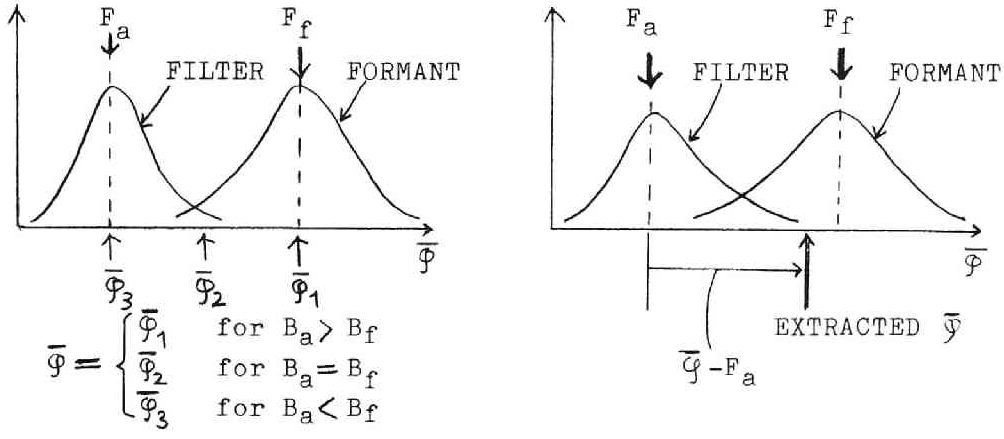
Fig 2.12 Zero-crossing distributions of formant signal passed through single tuned filters of different center frequencies. (Refer to Fig. 2.11.) Abscissa is frequency in cps. Ordinate is number of zero-crossings.

mant, we can estimate the formant frequency from $\overline{\varphi}$. The $\overline{\varphi}$ measured in actual experiment is only the approximation of F_f, when $F_f \neq F_a$. However, $|F_f - \overline{\varphi}|$ gives the information of the position of unknown formant F_f relative to the center frequency of filter F_a and polarity of $F_f - \overline{\varphi}$ gives the directon from F_a to F_f (See Fig. 2.13). The same principle can be applied for the zero-crossing distribution $W(f)$ and mean zero-crossing number $\overline{N}$.

The method to detect the formant frequency from phase characteristics such as the distribution of instantaneous frequency $p(\varphi')$, $W(f)$, $\overline{\varphi}$ and $\overline{N}$, is to prepare a bank of single tuned filters and to measure the phase characteristics of filter outputs. The block diagram is shown in Fig. 2.14. Each output of filters having center frequency $F_{a1}, F_{a2}, \cdots\cdots, F_{an}$ is converted to zero-crossing wave $x_1(t), x_2(t), \cdots, x_n(t)$. The principles of processing are: (i) $\overline{\varphi}$ or $\overline{N}$ is measured, which is compared with the standard value $\overline{\varphi}_o$ or $\overline{N}_o$ assigned for each filter. The measure of the closeness of the formant frequency F_f and filter center frequency F_a is given by $\overline{\varphi} - \varphi_o$ or $\overline{N} - N_o$, the minimum of which shows the formant. (ii) $p(\varphi')$ or $W(f)$ is measured for each channel, from which the components having the value φ' or f near to the center frequency of the filter is selected. The maximum output of these components shows the formant.

In this section discussion is made on method (ii) by measuring $W_i(f)$. The band width of filters must be $B_{ai} > B_f$ (B_{ai} is the band width of the i-th filter). The broader the B_{ai} the closer $\overline{\varphi}_i$ to F_f and $W_i(f)$ is distributed around F_f. On the contrary the broader B_{ai} lacks the ability to separate the adjacent formants. Taking into account these complementary condition, B_{ai} was chosen between 100 cps——300 cps.

The zero-crossing distributions obtained for the i-th filter of Fig. 2.14 is expressed as



(a) Ideal case (b) Experimental case ($B_a > B_f$)

Fig. 2.13 Estimation of formant frequency F_f by the instantaneous frequency $\bar{\varphi}$ extracted by passing the formant signal through single tuned filter located at F_a.

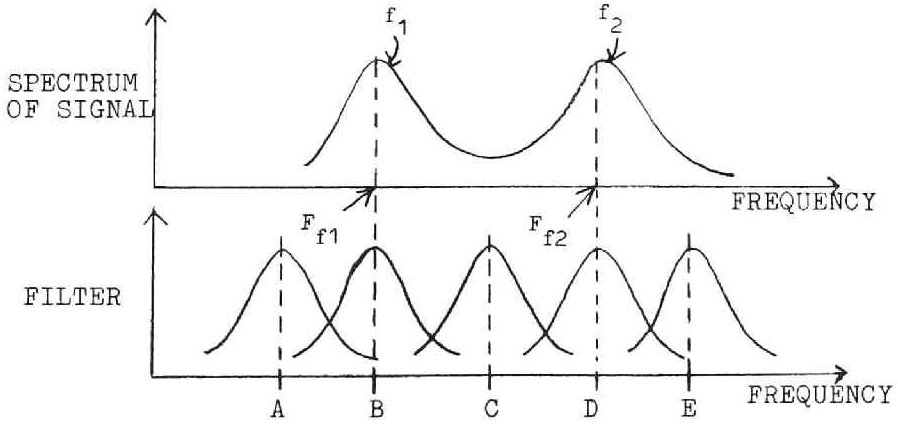
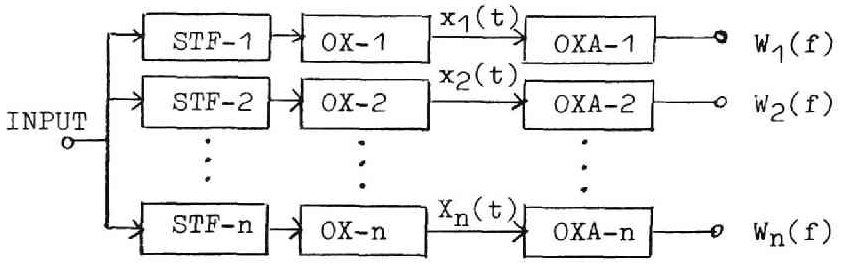
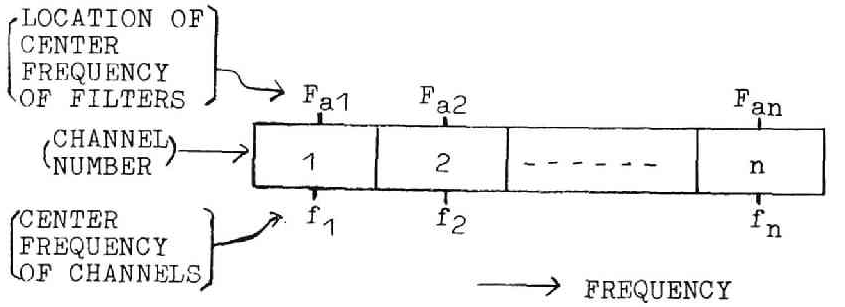


Fig. 2.15 Locations of single tuned filters to two formants.

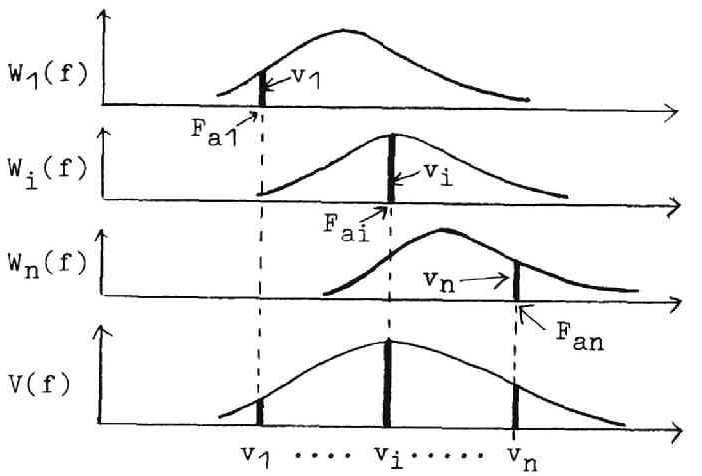


STF: SINGLE TUNED FILTER
OX: ZERO-CROSSING CONVERTER
OXA: ZERO-CROSSING ANALYZER

(a) Block diagram



(b) Frequency arrangement of the channel of $W_i(f)$.



(c) Transformation to V(f) from a set of zero-crossing distribution $\{W_i(f)\}$.

Fig. 2.14 Formant extraction by zero-crossing analysis using a bank of single tuned filters.

$$W_i(f) = w_{ij}(f) = w_i(f_j), \qquad \begin{matrix} i = 1, 2, \cdots\cdots, n \\ j = 1, 2, \cdots\cdots, n \end{matrix} \tag{2.21}$$

in which argument $f = \frac{1}{\tau}$ is used instead of the measured zero-crossing interval τ. The distribution $W_i(f)$ is expressed in the form of channel classified outputs, the number of which was coincided with the number of filters and the center frequency of each channel was chosen as $f_j = F_{aj}$ ($j = 1, 2, \cdots\cdots, n$) as shown in Fig. 2.14.(b).

The formant is detected by obtaining new distribution $V(f)$ from a set of $\{W_i(f)\}$ as shown in Fig. 2.14(c). Let's denote the channel output of $\{w_i(f_j)\}$ that satisfy $j = i$ as v_i ; that is

$$w_{ii} \quad v_i$$

The distribution given by

$$V(f) = \{v_i\}, \qquad i = 1, 2, \cdots\cdots, n \tag{2.22}$$

is called here "summed (zero-crossing) distribution." Then, the peak of $\{v_i\}$ will show the presence of formant and formant frequency is given by

$$F_f = F_{ap} \quad (p \,|\, v_p = \max\{v_i\}) \tag{2.23}$$

In the realization of circuits, we need not classify $\{w_{ij}\}$ for $i \neq j$, but have only to extract v_i from $x_i(t)$.

3. Zero-crossing Analysis of Signal with Two Formants

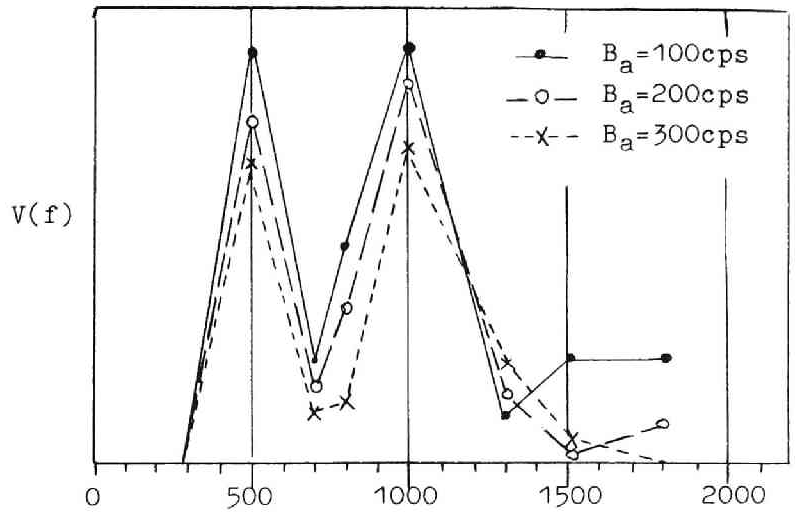
In the above discussion the signal to be analyzed was simplified to have only one formant. The speech signal, however, contains some formants. Since the mean spacing between formants is decided by the length of vocal tract, there is no occasion that the three formants gather closely. It is necessary to examine how the formant extraction scheme by the zero-crossing distribution using single tuned filter works on the signal of this type. As the problem is difficult to treat theoretically, some experiments was tried on synthesized signals and on speech sounds. Considering the limitation of the distribution

of formant frequencies and the characteristics of single tuned filter, we may discuss the signal with two formants.

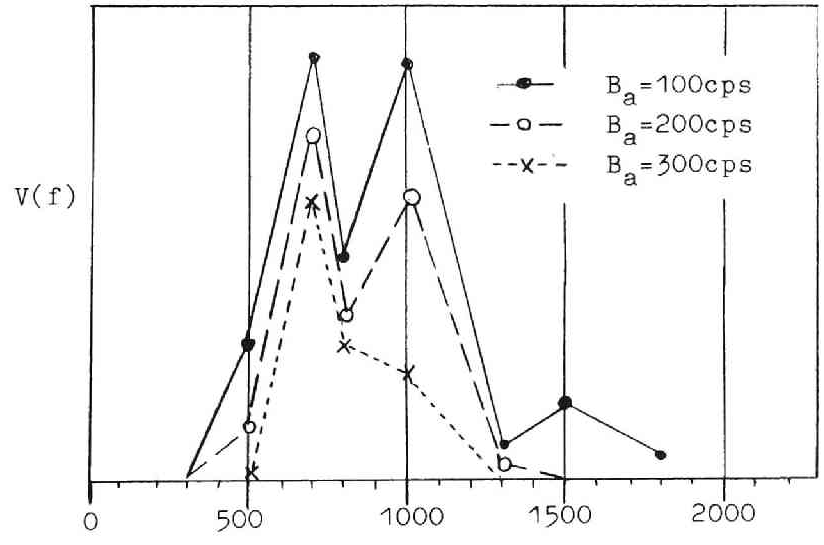
The relative locations of a filter to two formants situated at F_{f2} and F_{f1} are shown in Fig. 2.15. The contribution of formant f_2 (or f_1) to the response of filter located at A (or E) is slight compared with the contribution of f_1 (or f_2) . When filter location is matched to one of the formants as in cases B or D and the distance of formants f_1 and f_2 is not so large, considerable amount of disturbance willbe exerted from the other formant, the degree of which will be proportional to the band width of filter. This requirement is complementary with the requirement discussed above about one formant signal.

Another problem arises when filter is situated between the formants as in case C of Fig. 2.15. If the amplitudes of f_1 and f_2 are equal the signal itself has the mean frequency at $(F_{f1}+F_{f2})/2$, which can only exist under the unstable condition. When the signal is passed through the filter situated at C having narrower band width than formants, the resultant signal will possibly have mean frequency at $(F_{f1}+F_{f2})/2$ as stable condition. Therefore, the false formant might appear. If f_1 and f_2 are closely located, F_{f1}, $(F_{f1}+F_{f2})/2$ and F_{f2} will be detected as the one broad formant. To suppress this apparent formant, the wider band width of filter is recommended. In actual speech sound, the zero-crossing intervals are widely distributed around mean frequency so that the summed zero-crossing distribution $V(f)$ will not have appreciable output at this frequency.

Fig. 2.16 shows the summed zero-crossing distributions $V(f)$ of the signal with two formants at F_{f1} and F_{f2} passed through filters located at various frequencies. In (a), as the difference of both formant frequencies is 500 cps, the separation of both formants is clear for B_a =100, 200 and 300cps. On the contrary in (b), the spacing of formant frequencies is 300 cps so that the wider band width of filter causes the interaction between both formants.



(a) F_{f1}=500cps F_{f2}=1000cps
B_{f1}=100cps B_{f2}=100cps



(b) F_{f1}=700cps F_{f2}=1000cps
B_{f1}=100cps B_{f2}=100cps

Fig. 2.16 Summed zero-crossing distibution V(f) of the two formant signal for the various band widths of single tuned filter B_a. The formants are located at F_{f1} and F_{f2}, having the band width B_{f1} and B_{f2}, respectively. Abscissa is frequency in cps. Ordinate is V(f).

For B_a=300 cps the second formants at F_{f2}=1000 cps is not detected.

4. Application of the Method to Speech Sound Analysis

The zero-crossing analysis using single tuned filter described in previous section was applied to the analysis of vowels and some consonants. The method is essentially indifferent to the range of the formant frequencies of male and female voice, although additional circuit devices are needed than in zero-crossing analysis by band pass filters. The procedure is the same as in Fig. 2.14. The results are shown in Fig. 2.17——Fig. 2.19.

(1) Vowel sounds and stop consonants.

From the result of Fig. 2.16, the band width of B_a =200 cps was used. The initial section of the vowel sound was sampled, as well as the middle part, to compare the distribution with that of the burst of plosive unvoiced consonant. The distributions of vowels are shown in Fig. 2.17(a)——(f) and consonants in Fig. 2.18(a)—— (f).

It is seen from these data that the first formant F_1 and the second formant F_2 of vowel are clearly seen even at the initial excitation, although the formants higher than 2 kc are not always evident. The adjacent formants F_1 and F_2 of /a/, /o/ are separated to some extent. The distribution of middle part of speaker S (Fig. 2.17(e) and (f)) can be compared with spectrum patterns or sections of APPENDIX I. From the spectrum sections sketched in Fig. 2.20, it is seen that F_1 and F_2 of material S-/a/ is fused into one peak at about 1——1.6 kc and the interformant components between F_2 and F_3 does not show remarkable valey, which description will be true in the distribution of S-/a/ of Fig. 2.17(e). For the other sounds, it is also found that the zero-crossing distribution of vowel has the similar tendency to its spectrum.

The results of stop consonants /p/ and /t/ are shown in Fig. 2.18(a)—— (f) for speakers S and D. The sampling was made in the initial 10——20 ms after the burst. The patterns do not show remarkable formant structure as in

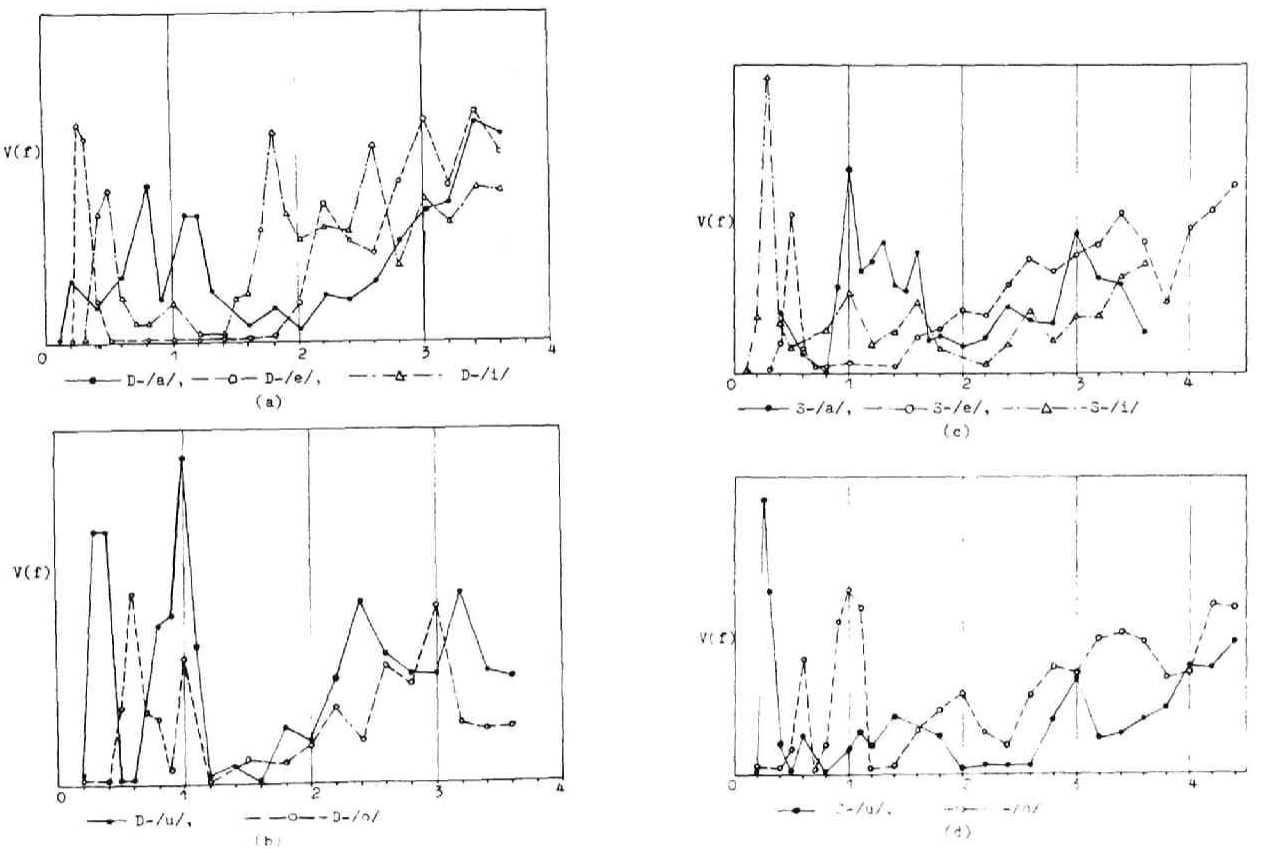
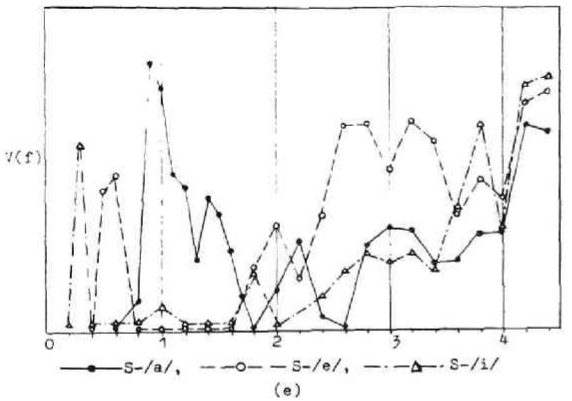


Fig. 2.17 Summed zero-crossing distributions $V(f)$ of vowel sounds sampled at the initial and the middle part, for speaker D(male) and S(female). The band width of the single tuned filter is set to B_f=200 cps. Abscissa is frequency in kc. (a)and (b): for speaker D(sampled at initial), (c) and (d): for speaker S(sampled at initial), (e) and (f)on next page: for speaker S(sampled at middle).



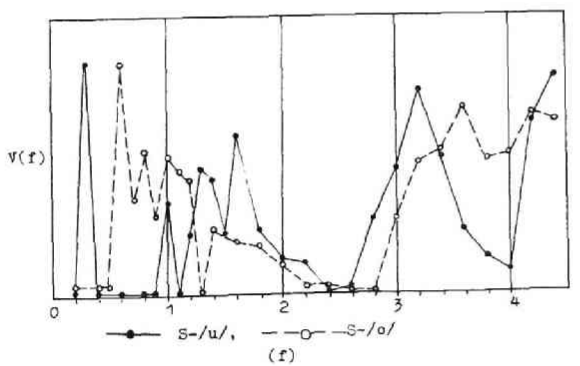
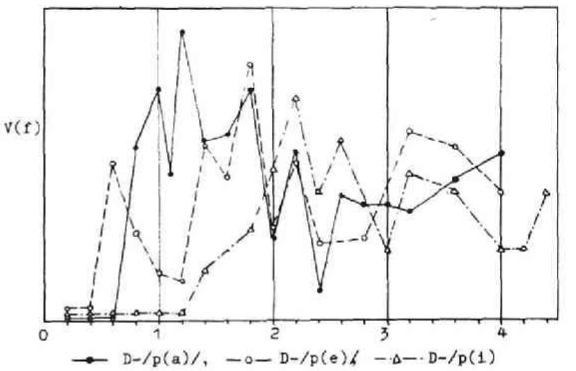


Fig. 2.17



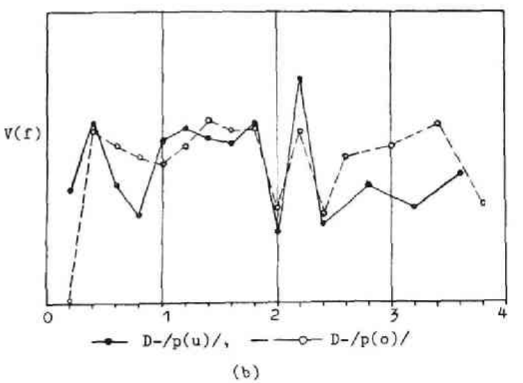
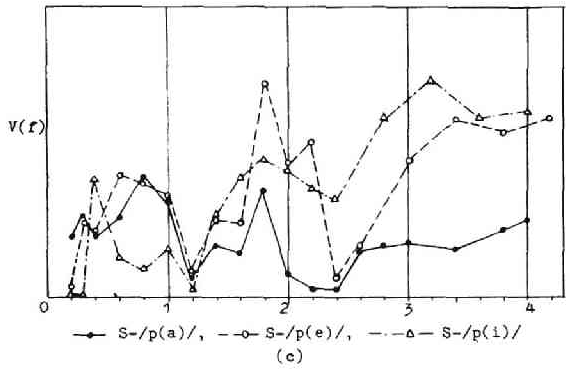
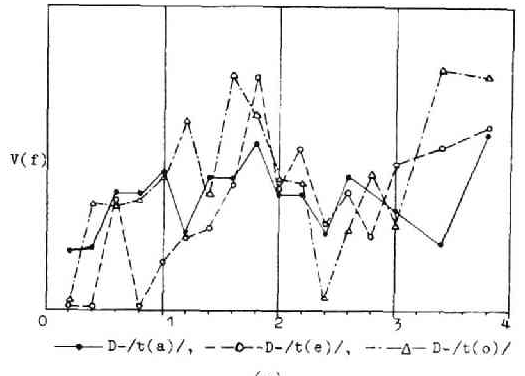
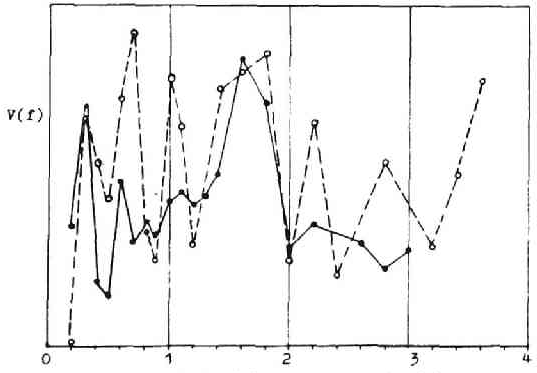


Fig. 2.18







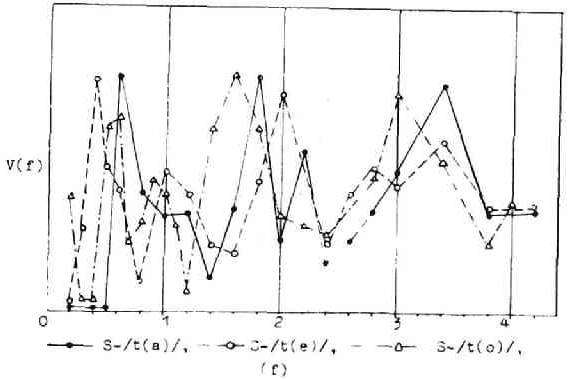


Fig. 2.18 Summed zero-crossing distributions V(f) of plosive consonants . The band width of the single tuned filter is set to B_a=200cps. Abscissa is frequency in kc. (a) and (b); /p/ of speaker S, (c) and (d); /p/ of speaker S, (e); /t/ of speaker S, (f); /t/ of speaker S. (a) and (b) are shown on previous page.

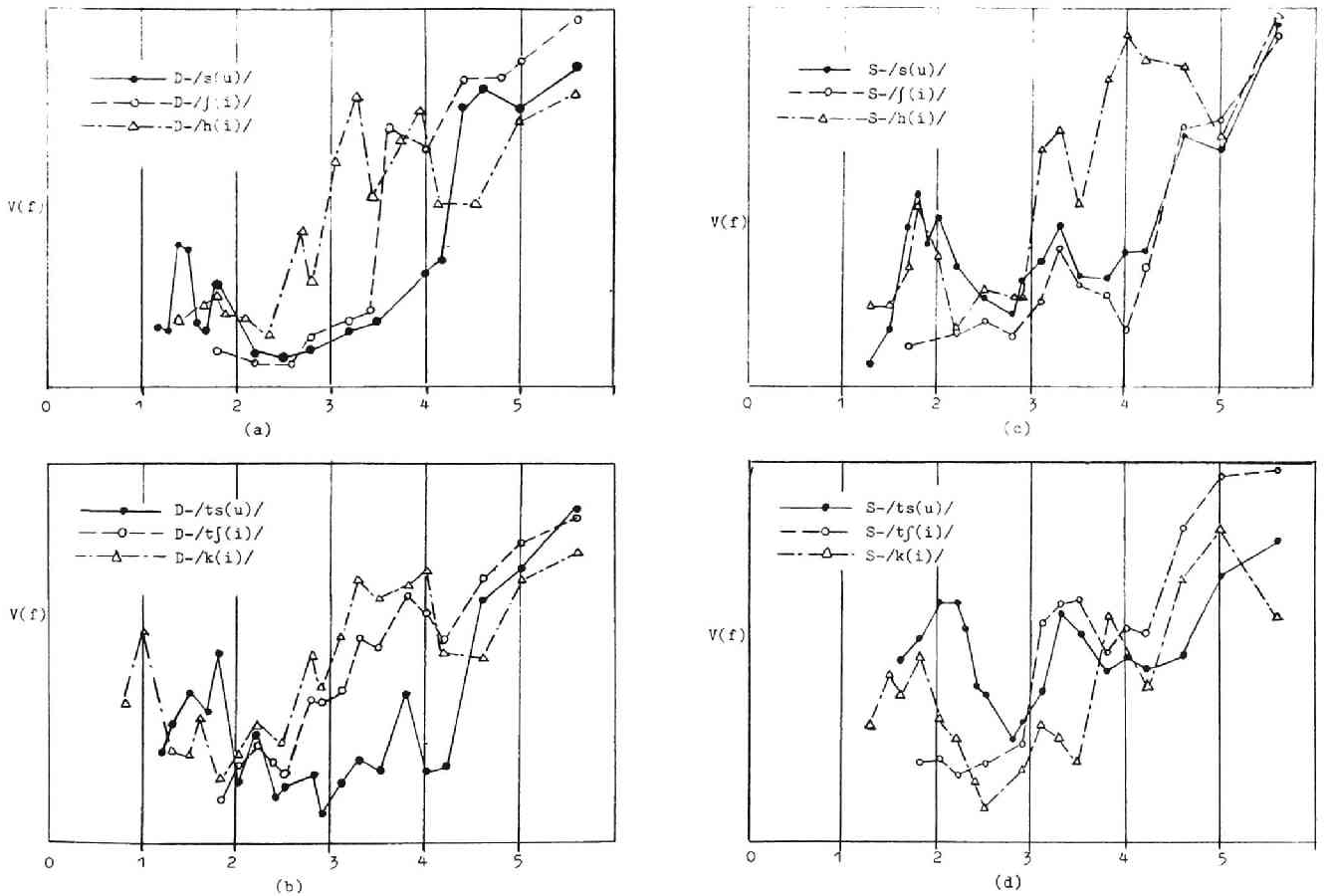


Fig. 2.19 Summed zero-crossing distrinutions V(f) of noise consonants sampled during 50ms at the last part of sound. Band width of single tuned filter was set to B_a=100cps. Abscissa is frequency in kc. (a) and (b); for speaker D, (c) and (d); for speaker S, (e) , /s/, /ʃ/ and /h/ in connected speech in the context of vowel elision, in which the underlined sounds were analyzed. (e) is shown on next page.

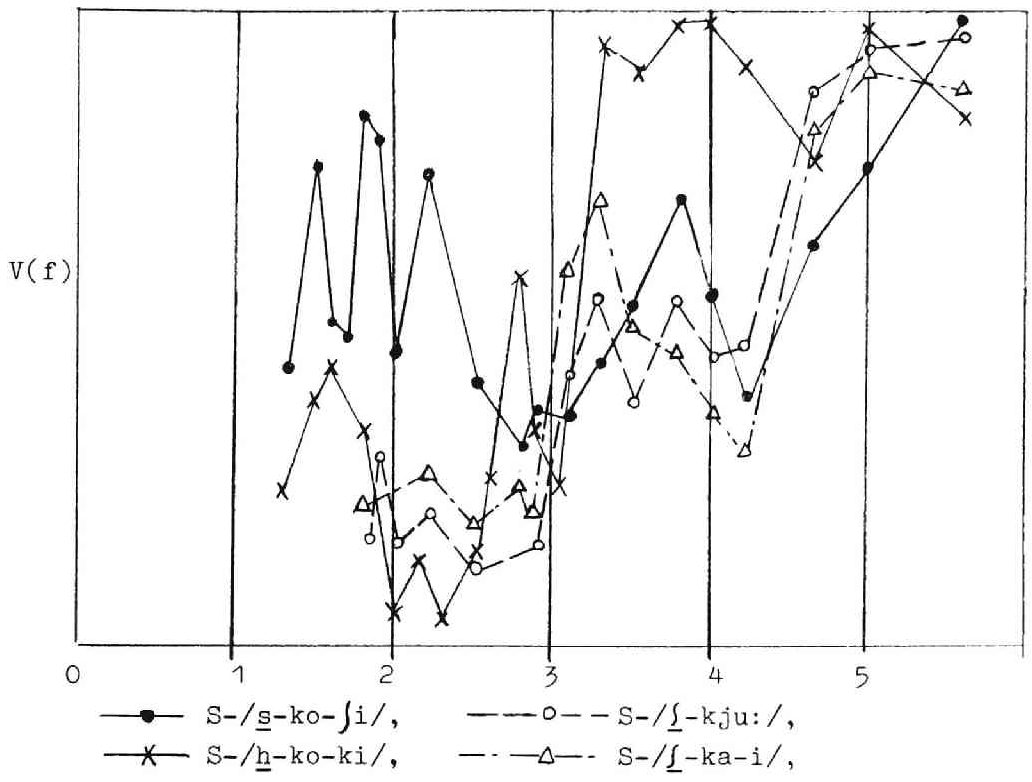


Fig. 2. 19(e)

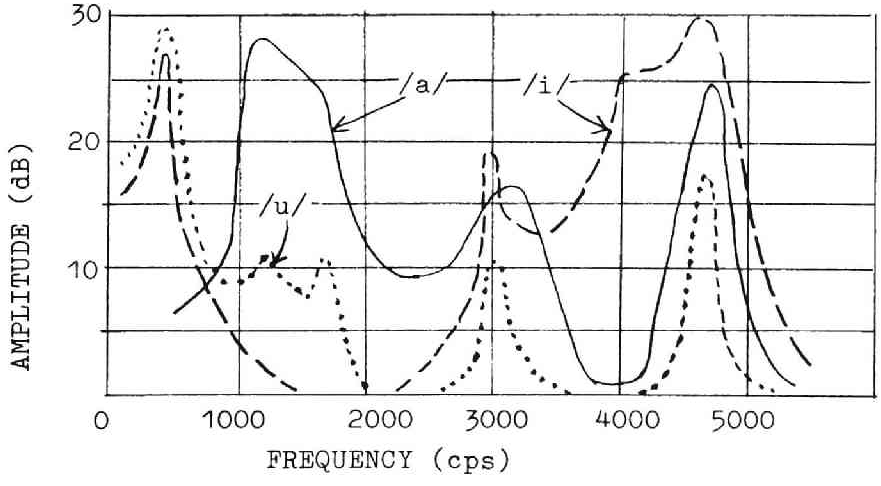


Fig. 2.20 Sketch of spectrum sections of the middle part of /a/(solid line), /i/(broken line)and /u/(dotted line) of speaker S. The materials are the same as used in Fig. 2.17.

the initial part of vowel, but distributed more uniformly. Some dominant peaks are observed in S-/t/ of Fig. 2.18(f), but the peak positions differ from that of vowels and the valey of distribution is not deep.

(2) Noise consonant

As discussed in chapter 4 of PART I, spectrum of fricative consonants (including /h(i)/) and affricate consonants (including /k(i)/) of Japanese have not shown distinct differences common to several speakers as the major formants or peaks are concerned. They suffer from the influence of speakers greatly. Although the lower formants are heavily damped and their levels are low, they may be the important cues of discrimination.

The zero-crossing wave is less affected by the level of signal than spectrum. The formant detection of F_2 and F_3 of noise consonants was tried by zero-crossing analysis using single tuned filters stated above. The 50 ms length was sampled of the last section of consonant sound. To lessen the disturbance of the intense, higher formants to the weak, lower formants, the band width of single tuned filter B_a =100 cps was used. The results are shown in Fig. 2.19(a)——(e). For speaker D, the lower end of cutoff of dominant formant is clearly different. On the contrary for speaker S(female), the cutoff of /s(u)/ and /ʃ(i)/ are almost the same (at about 4.5 kc). As the major formants of speaker S is located at higher frequency, the lower formants F_2 and F_3 are separated from these, but for speaker D, the F_3 seems to be masked by the major formants. The appearance of lower formants of speaker S is: /s(u)/ and /ts/ have F_2 at about 2 kc and F_3 at about 3.5 kc, /ʃ(i)/ and /tʃ/ have F_3 at about 3.5 kc, /h(i)/ and /k(i)/ have F_2 at about 2 kc and F_3 at about 3 - 4 kc.

In Fig. 2.19(e) the materials were selected from connected sound in which the noise sound of interest is not followed by vowel and therefore the distinction between /s/, /ʃ/ and /h/ must be made without the knowledge of the following vowel or of the transition to it. The above description on the lower formants of speaker S is valid for these materials, too.

2.5 Conclusion

In this chapter the descriptions were made on the representation of the zero-crossing wave, the analysis circuit of zero-crossing intervals and zero-crossing analysis of formant signals and speech sounds. A notion of phase sampling was proposed, by which the zero-crossing wave is considered as the sampling of the signal at every π rad. It was shown that the improvement of the articulation score of zero-crossing wave by differentiation has relation with the spectrum of the zero-crossing wave. Zero-crossing analysis circuit was devised, which measures the zero-crossing intervals by digital method and obtains the zero-crossing distribution expressed by pulse numbers and then expressed by a set of voltages, operating successively under the control of periodic sampling signal. This circuit was used as the snalysis circuit in the speech recognition system of chapter 3 and chapter 4.

Results of zero-crossing analysis suffer much influence from the characteristics of filter used before the conversion to zero-crossing wave. In this chapter the analysis of formant structure was performed by passing the signal through single tuned filter. The processing on the envelope of the signal passed through single tuned filter was tried in chapter 2 of PART I and the phase characteristics were taken in this chapter. The sequence of zero-crossing intervals of the signal with one formant, passed through single tuned filter, was found to vary largely according to the relation of band widths of both resonant circuits. Some methods were proposed that extract the formant structure of the speech sound by the analysis of zero-crossing waves passed through single tuned filters. The method was adopted that obtains so called "summed zero-crossing distribution" from a set of the zero-crossing distributions stated above. The formant structure at the onset of vowel was compared with that of the burst of unvoiced consonant, by which the differences of those signals were presented. The method was also applied to the analysis of noise sounds having stationary noise at high frequency, such

as /s/, /ʃ/, /h(i)/, /ts/, /tʃ/ and /k(i)/, and could detect the lower formants (the second or the third) which are often masked by higher frequency components. It was deduced from the results that the behaviors of such formants are important in the distinction of the noise consonants.

Chapter 3

SPEECH RECOGNITION SYSTEM OF JAPANESE SOUNDS (24) (9)

3.1 Introduction

In this chapter the automatic recognition system of Japanese speech sounds is described.

As has been stated in chapter 5 of PART I, the processing steps of speech recognition are divided into several levels such as: (1) analysis and parameter extraction in acoustic level, (2) processing in phonetic level and then (3)processing in linguistic level, although the processings in these levels must be co-operatively performed.

The recognition system of this chapter processes the speech sound from the acoustic and the phonetic stand point of view. One of the principal problems of the speech recognition system is the selection of the recognition unit. There have been several trials of the recognition system of limited vocabulary (for instance, spoken digit recognizer), in which the word (or a whole speech sound wave continuously spoken) was treated as a whole without separating the speech wave into the smaller recognition units that constitute the wave. This method may, however, be applicable only for the limited vocabulary where its number is not so large. For the recognition system that accepts the general vocabulary not limited to a certain particular words, the word must be decomposed into smaller units. Here the phoneme was selected as the basic recognition unit.

The system stated here can accept a short words as well as monosyllables of Japanese. It has two functions; one is the segmentation of the speech sound wave into several segments corresponding to the phoneme and the other is the analysis and recognition. In segmentation the speech sound is separated into consonant sections and vowel sections and then the vowel section is divided into several segments of vowel phonemes. In the recognition part there

prepared several circuits, each designed for the parameter extraction. The speech sound is classified into several categories and the consonant sections and the vowel sections are separately analyzed. The segmentation part is combined with recognition part at the final stage to control the timing of output.

The equipment is transistorized and is composed of the unit packages. The registers, memories and shift registers are composed of static flip-flop circuits, and the logics of phoneme classification, zero-crossing interval classifier and phoneme recognition circuit, etc., are performed by diode logical circuits. For a complicated decoding logic, the programmable matrix board is used by means of a plug-in diode method. The shift register operates as a working register for matching operation as well as a temporary memory space for the input pattern. A size of the machine, since it is an experimental model, is fairly large to make it easy to modify and add circuits.

3.2 Principle of Speech Recognition System

Fig. 3.1 shows the whole block diagram of the conversational speech recognition system. It is divided into two parts according to the function; the segmentation part (I of Fig. 1) which separates the input speech sound into discrete sections and the recognition part (II of Fig. 1) which performs the discrimination of the separated sections.

1. Segmentation Part

The principal operations of the segmentation part are to distinguish sections(segments) corresponding to the recognition unit (phoneme) from the time pattern of parameters extracted from the input speech sound and thereby to control the operation of the recognition part, such as sampling, discrimination and output timing.

From the pattern obtained by Sonagraph, the observed change of the speech

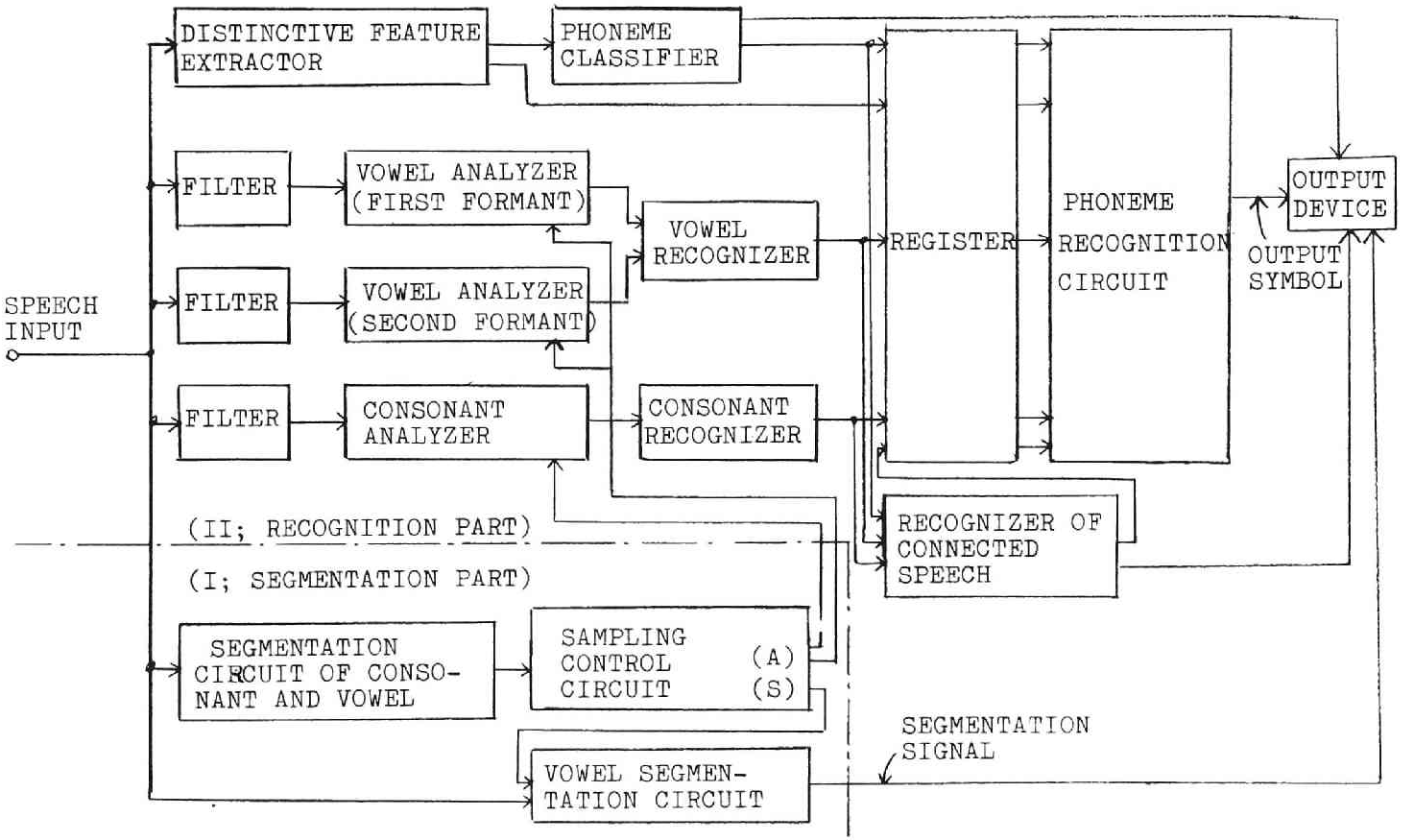


Fig. 3.1 Block diagram of speech recognition system.

parameters in the vowel section is gradual. Therefore it is impossible to define a clear separating point of the sections in the vowel section, whereas we can separate the speech sound into such sections as vowel-like section noise-like section, buzz-bar section.

In segmentation part the "segmentation circuit of consonant and vowel" divides the speech sound into consonant sections and vowel sections. Further, when vowel section contains more than one recognition unit (phoneme), the "vowel segmentation circuit" divides the vowel section into segments each of which contains the section of the speech sound wave corresponding to the vowel phoneme. For this operation of segmentation the pattern of time variation of speech parameters is examined. This is based on the fact that speech sound is composed of such sections as; quasistable section in which the parameters remain almost constant state, transient section in which the parameters move gradually and the section or time point at which the parameters make abrupt change. The principle of the segmentation between consonant section and vowel section is similar to the phoneme classification in the recognition part.

Segmentation in vowel section is performed by extracting the parameters, that describe the aforementioned sections, from the digital pattern of speech sound analyzed by the zero-crossing wave analysis. As the parameters we defined two quantities called "distance" and "stability". Stability expresses the stationary property of pattern, that is, the property that the parameter remains almost constant state over a certain period. On the contrary, distance is the quantity that expresses the change of the pattern. Stability is useful to find out the vowel and the fricative and distance is to the sound with burst such as stop consonant.

In the operation of segmentation, the selection of the recognition unit is the basic problem. To accept the conversational speech sound we selected the phoneme rather than the morpheme or the word, and the phonetic context was treated in the another part of the machine. (Phonetic context is discussed in the next chapter.)

2. Recognition Part

The recognition part performs the recognition of the speech segment, segmented by the signal from the segmentation part, and outputs the phoneme. In phoneme classifier of Fig. 3.1 speech wave is classified to the several groups, each of which corresponds to the manner of articulation in our speech organ, by the distinctive features. In parallel with this operation vowel recognition system and consonant recognition system, each of which is composed of filter, analyzer and recognizer, find the parameters corresponding to the place of articulation and recognize phonemes which are classified as belonging to the same group by the phoneme classifier. The analyzing method is the zero-crossing wave analysis which is applied to the consonant section and vowel section separately, whose samplings are controlled by the sampling signal from the sampling control circuit. Sampling is only once for one consonant section and for vowel section it is periodically repeated. All the results are once memorized in registers and combined in phoneme recognition circuit, the operation of which is controlled by segmentation signal, to send the recognized output symbol to the output device.

As mentioned above we selected as recognition unit the phoneme and the phonetic interaction between phonemes which is essential to the conversational speech sound is treated in the recognizer of connected speech of Fig. 3.1. All the results from the previous stages of the machine are once stored in the register and are combined in the phoneme recognition circuit where final recognition of phoneme is made, controlled by the segmentation signal from the segmentation part. The output symbol is the Kana letter, the Japanese phonetic alphabet and also the orthography.

3.3 Segmentation to Recognition Unit [31][32]

1. Distance and Stability

As speech sound is analog and continuous signal, its recognition is after

all the coding of the speech signal to the letter symbol. Further, it will be desirable to process the speech sound in digital form. Here the input speech wave is converted into the digital pattern which is the time series of parameters by digitizing the analog parameters with appropriate unit and by sampling it with the time unit sufficient to maintain the characteristics of time variation of the speech sound.

The information of speech sound wave is, according to the sampling theory, too large to treat it directly but on the other hand it contains much redunduncy when the linguistic information of speech sound is considered. From the view point of time domain, typical sections of the speech sound are; a) quasi-stable state in which parameters at each time point are closely related with each other and are repeated only with slight change, b) transitional section in which parameters change is gradual except some time points at which parameters change abruptly. Segmentation may be performed by paying attention to such time change characteristics of parameters. For this purpose two criteria "stability" and "distance" were defined as follows.

On the time axis the time points 1, 2, ... j, j+1, ... are selected and the interval between j-th and j+1-th time points is called j-th sampling interval. Let's denote P_{ij} as the i-th element of parameters or distributions (i=1, 2, ... , n) in the j-th sampling time interval, normalized in each time interval. Then

$$P_j = \{P_{1j}, P_{2j}, \cdots\cdots, P_{nj}\} \tag{3.1}$$

is the parameter set or distribution in the j-th time interval. And the whole pattern of speech sound is expressed as follows;

$$P = \{P_j\} = P_1, P_2, \cdots\cdots, P_j, \cdots\cdots \tag{3.2}$$

We define the index of stability $X_{ij}(\ell)$ as

$$X_{ij}(\ell) \quad \frac{1}{\ell}\sum_{k=0}^{\ell-1} P_{i\ j-k} \tag{3.3}$$

where ℓ is the number of sampling intervals. $X_{ij}(\ell)$ can be defined in each time interval and in each channel and has the value 0-1. This index represents the rate in the i-th channel during the ℓ intervals before the j-th sampling interval and represents the stationarity of pattern near the j-th time interval. When this takes a large value it may be considered that the neighbouring part of the pattern belongs to one segment corresponding to a phoneme. Thus by selecting a proper value of ℓ, $X_{ij}(\ell)$ gives an important information in distinguishing the stationary part of pattern from the transient part which inevitably appears between the stationary parts.

The distance d_j is defined by

$$d_j = \sum_i |P_{ij} - P_{i\ j-1}|$$

where the sign $\sum$ may be rewritten by $\oplus$ (exclusive OR), when P_{ij} is a variable of Boolean algebra and d_j is then the Hamming distance. The distance is the quantity to measure the magnitude of change of pattern and takes a large value for the time interval where pattern makes abrupt change and a small value for the stationary interval where the pattern keeps a nearly constant state.

2. Segmentation of Successive Vowels by Zero-crossing Wave Analysis

Experiment of segmentation using the zero-crossing pattern was applied to the vowel segmentation circuit of Fig. 3.1. Fig. 3.2 shows the block diagram of this circuit. The zero-crossing distribution

$$\{W_{ij}\} \qquad i = 1, 2, \ldots\ldots, n$$

is obtained in each of successive sampling interval as have been explained in chapter 2, section 3. Therefore the pattern is formed as the series of zero-crossing wave distributions.

The analysis of speech sound during, for example, 10 ms gives channel classified distribution $\{W_{ij}\}$. Then W_{ij} is digitized to aforementioned P_{ij} by setting up a threshold level relative to the maximum value at that sampling interval. The processing of pattern for distance and stability is made in

shift register having the length of ℓ. Though this method may be applied to the general speech sound, we applied this to the vowel section by the reasons that, because of the "consonant + vowel" construction of the Japanese syllable, vowel has important function and that the scale of hardware is too large to apply the method to the whole speech sound.

As the parameter of vowel, the first formant and the second formant were selected. The circuits used to obtain zero-crossing pattern $\{W_{ij}\}$ in Fig. 3.2 is identical to the analysis circuit explained in chapter 2. Before the zero-crossing wave analysis, frequency regions of the first formant (F_1) and the second formant (F_2) are picked up by passing the input speech sound through the filters, low pass filter of 1,500 cps for F_1 region and band pass filter of 800 - 2,500 cps for F_2 region. (The first formant and the second formant are expressed with the suffix (1), (2) respectively, but the operation of both regions being the same as is seen in Fig. 3.2, its suffix is often omitted.) In Fig. 3.2 zero-crossing wave converter (O-X converter) converts input speech wave into zero-crossing wave, a series of rectangular waveforms in which only the time points at which the original speech sound crosses zero level are left as the information bearing parameters and the wave has a constant positive or negative level according to the polarity of the original wave.

The zero-crossing wave analysis[(1)(2)] is executed in the zero-crossing interval classifier (O-X classifier) of Fig. 3.2, and the classified number of width is integrated as zero-crossing wave distribution for a certain sampling interval T. By repeating the sampling in a constant period, the zero-crossing pattern is obtained.

A statistical expression of zero-crossing interval measurement under the successive sampling is

$$W_{ij} = W_j(\tau_i) = \frac{1}{T}\frac{N_{ij}\tau_i}{\Delta\tau_i} \qquad \begin{matrix}(i=1, 2, \cdots\cdots, n)\\ (j=1, 2, \cdots\cdots\cdots)\end{matrix} \qquad (3.5)$$

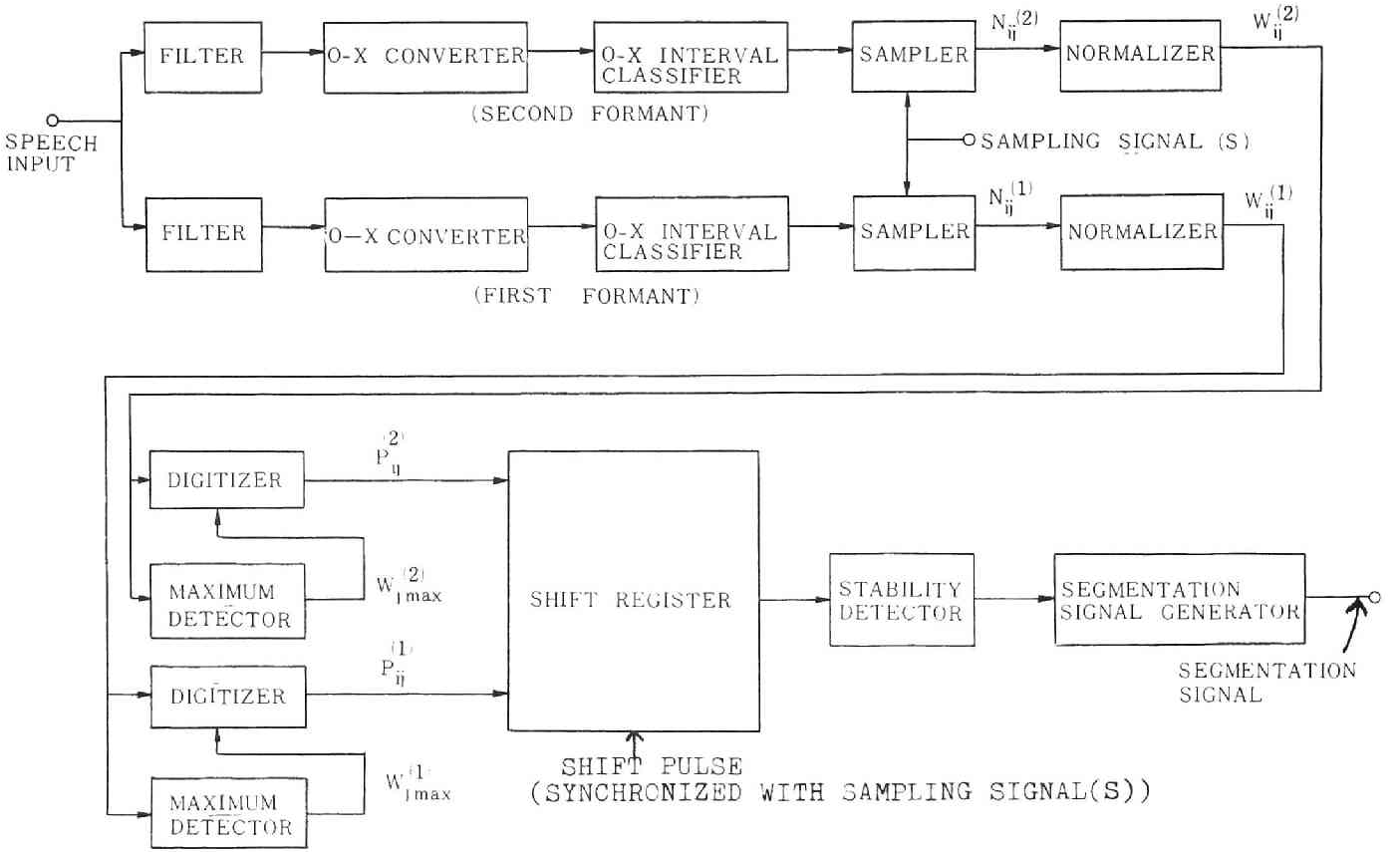


Fig. 3.2 Block diagram of vowel segmentation circuit.

and the zero-crossing distribution in the j-th interval is

$$W_j = \{W_{1j}, W_{2j}, \cdots\cdots, W_{nj}\}$$

(suffix n may be n_1 for F_1 and n_2 for F_2). By the multiplication of τ_i as seen in the above expression (3.5), W_{ij} represents the ratio of the time intervals, which are the summation of the widths of the rectangular waves clasfied to the i-th channel during the sampling interval T, to the total time T. Thus $\sum_{i=1}^{n} W_{ij}$ has almost constant value.

The circuit of zero-crossing wave analysis has the same construction with the vowel analyzer. Therefore as shown in Fig. 3.10 and Fig. 3.11, a pair of the zero-crossing distributions in F_1 region and in F_2 region, each of which having n_1 and n_2 channels, respectively, is obtained.

The normalizer of Fig. 3.2 converts (like the integrating counter of Fig. 2.7) the N_{ij}, which is given as pulse number, to the proportional analog voltage W_{ij} of the above equation. By separating the components of a formant from the others by using filter, the zero-crossing patterns are closely related with the spectra of orignal speech sound and its peak corresponds to the formant. Therefore we can obtain simply and effectively the formants by the following method. Maximum detector of Fig. 3.2 detects the peak value $W_{j\,max}$ of the n channel distribution in each time interval and digitizer converts W_{ij} to the aforementioned P_{ij} with the threshold level $W_{j\,max}/\alpha$ ($\alpha \geqslant 1$) (When $\alpha = 1$, the digitizer operates as peak extractor.). This digitized pattern P_{ij} is sent to the shift register of Fig. 3.2.

Fig. 3.3 shows an example of the digitized pattern. The distribution $P_{ij}^{(1)}$ for F_1 region and $P_{ij}^{(2)}$ for F_2 region, which have 5 and 8 channels respectively, are separately digitized with the threshold level $\alpha \fallingdotseq 1$. In the figures the sampling is repeated every 10 ms (T = 10 ms) and one black point represents 10 ms. In the upper part of the pattern a series of vowel recognitions repeated every 20 ms is added. (As the pattern display cycle is 10 ms, one recognition of vowel series is shown by a succession of two black

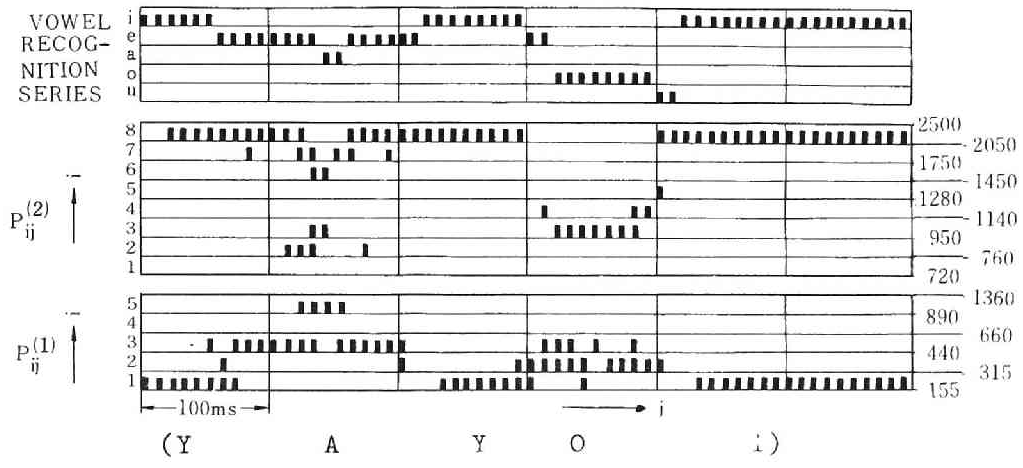


Fig. 3.3(a)Digitized pattern of zero-crossing distribution and a series of vowel recognitions for input sound "YAYOI" with the channel classification characteristics(cps).

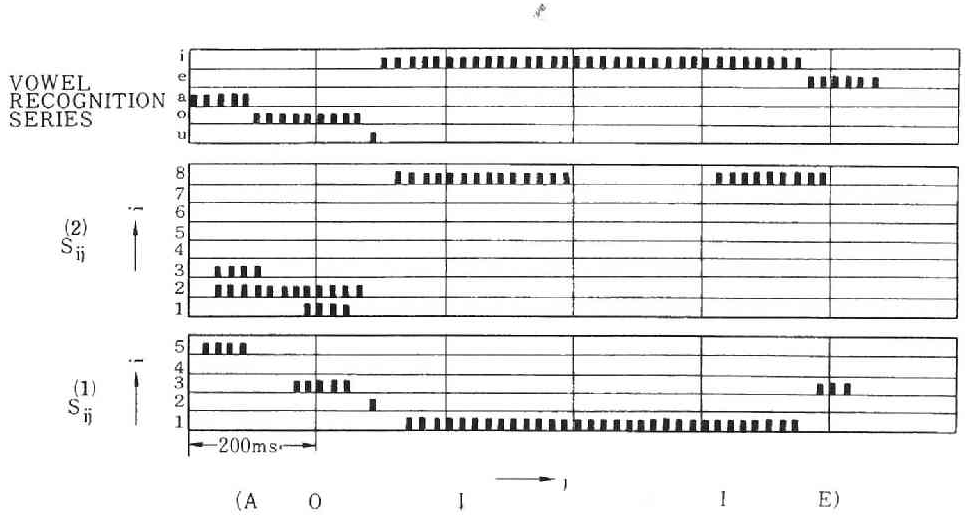


Fig. 3.4 Stability pattern and vowel recognition series of a connected vowel sound "AOI-IE"(Japanese). The channel arrangement is the same as that of Fig. 4.

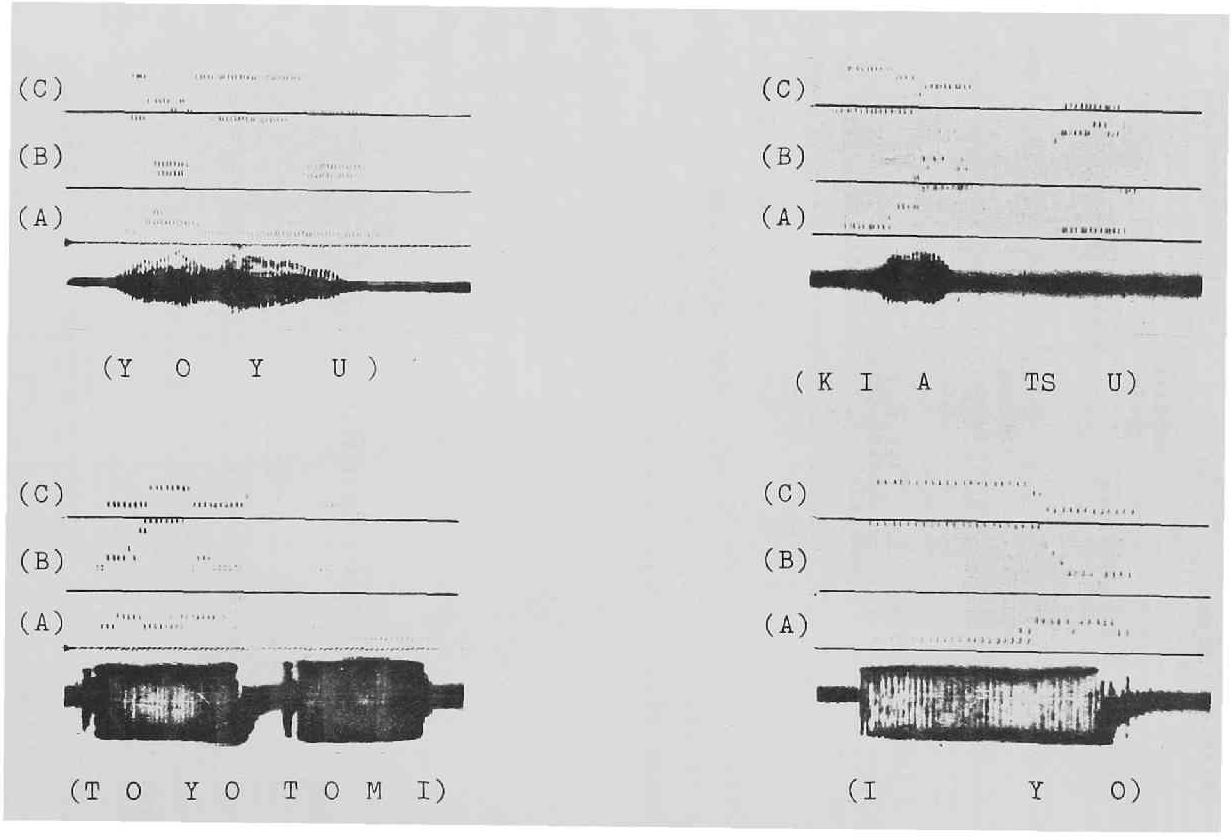


Fig. 3.3(b) Examples of photographic data of digitized pattern of zero-crossing distribution. (Refer to Fig. 3.3(a).)
(A); $P_{ij}^{(1)}$, (B); $P_{ij}^{(2)}$, (C); Vowel recognition series.

points.) The shift register of Fig. 3.2 memorizes the digitized pattern for the required time for processing, by which the time pattern is expressed in the form of space instead of time. The register has 13 channels and ℓ bit length enough to detect the stability.

The stability detector computes the stability $S_{ij}(^{h}/_{\ell})$ by digitizing the index of stability taken from the pattern $\{P_{ij}\}$ in the shift register.

$$S_{ij}(^{h}/_{\ell}) = 1, \quad \text{when} \quad X_{ij}(\ell) \geqslant {}^{h}/_{\ell}$$

$$S_{ij}(^{h}/_{\ell}) = 0, \quad \text{when} \quad X_{ij}(\ell) < {}^{h}/_{\ell}$$

For the vowel interval, $S_{ij}(^{6}/6)$ and $S_{ij}(^{4}/5)$ were used and appropriate one of them was selected for each channel. An illustration of the stability is shown below.

Sampling Interval (j)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Input Pattern (i)	0	1	1	1	1	1	1	1	0	1	0	1	1	1	0	0
$S_{ij}(^{6}/6)$	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
$S_{ij}(^{4}/5)$	.	.	.	.	1	1	1	1	1	1	0	0	0	1	0	0

As seen in the above example, $S_{ij}(^{4}/5) = 1$ means that in i-th channel there appeared more than four "1" during the five successive intervals just before the j-th sampling interval. From its definition, a stability pattern $\{S_{ij}\}$ will be derived from the digitized pattern. An example of stability pattern is shown in Fig. 3.4, where α is set close to 1 and $^{h}/_{\ell}$ to $^{6}/6$. The pattern shows that noisy components are smoothed out and dominant channels are left. This is based on the averaging function of the stability detection. As the existence of the stability implies the existence of the formant, we recognize the section, during which the stability has been detected in both F_1 and F_2 regions, as one segment corresponding to a phoneme. (No operation is performed to find the clear time points separating two segments.)

The segmentation signal generator of Fig. 3.2 generates the segmentation signal every time when a new combination of the stability is detected, and this

signal controls the recognition part of Fig. 3.1 with the rule shown in Fig. 3.18.

The function of stability detection is influenced by the characteristics of the channel classification of the zero-crossing wave analysis, by the setting of the threshold value α of the digitizer and by the value ℓ. These values are determined by the experimental data so as to satisfy both the detection of the stationary part and the suppression of the transition part.

In Fig. 3.5 schematic diagram of the stability detector and the segmentation signal generator is shown.

The length of shift register on time axis is ℓ=6, therefore in the j-th time interval the stored pattern is as follows;

First formant pattern $P_{ij}^{(1)}$ i = 1, 2, ..., n_1 ; j=j, j-1,...,j-(ℓ-1)

Second formant pattern $P_{ij}^{(2)}$ i 1, 2, ..., n_2 ; j=j, j-1,...,j-(ℓ-1)

The contents of shift register in j-th time interval are shifted by one bit along the time axis at the next time interval. The logical circuits are connected to each memory cell of shift register. As all the circuits are static logic using transistors and diodes, the contents of shift register and the state of the logical circuits are held constant during one interval. The stability detectors $STD_i^{(1)}$ and $STD_i^{(2)}$ connected to each channel give $S_{ij}(^h/_\ell)$. $^6/6$ means the circuit for $S_{ij}(^6/6)$ and $^4/5$ for $S_{ij}(^4/5)$, the suitable one of which is selected in each channel by switch $SW_i^{(1)}$ and $SW_i^{(2)}$. Memory circuit $R_{i\ j-1}^{(1)}$ and $R_{i\ j-1}^{(2)}$ which memorizes the detected S_{ij} for one time interval and AND circuit $G_i^{(1)}$ and $G_i^{(2)}$ are used for the detection of the beginning point of the stability.

The outputs of the OR circuits $U_1^{(1)}$, $U_2^{(1)}$, $U_1^{(2)}$ and $U_2^{(2)}$ are as follows;

a_j ; shows the existence of the stability in F_1 region.

b_j ; shows the beginning point of the stability in F_1 region.

c_j ; shows the existence of the stability in F_2 region.

f_j ; shows the beginning point of the stability in F_2 region.

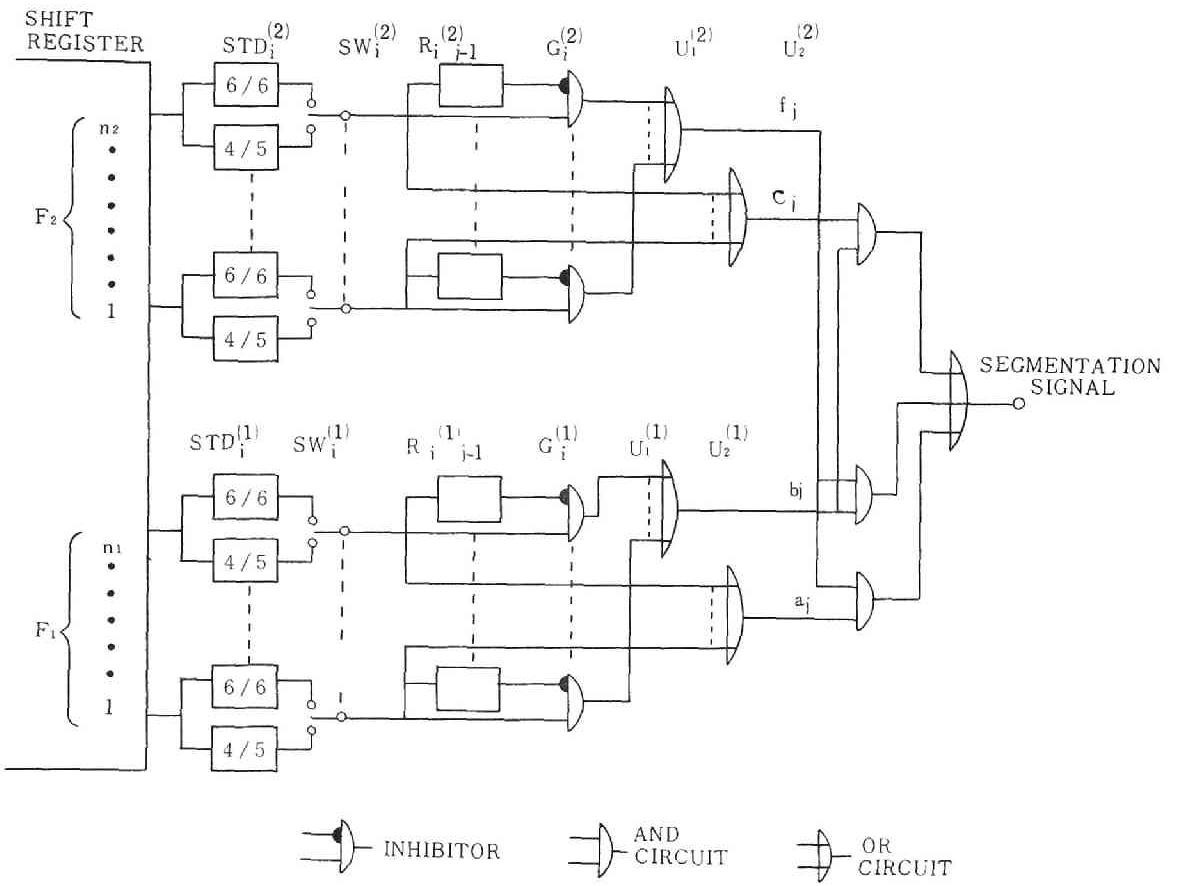


Fig. 3.5 Schematic diagram of stability detector and segmentation signal generator.

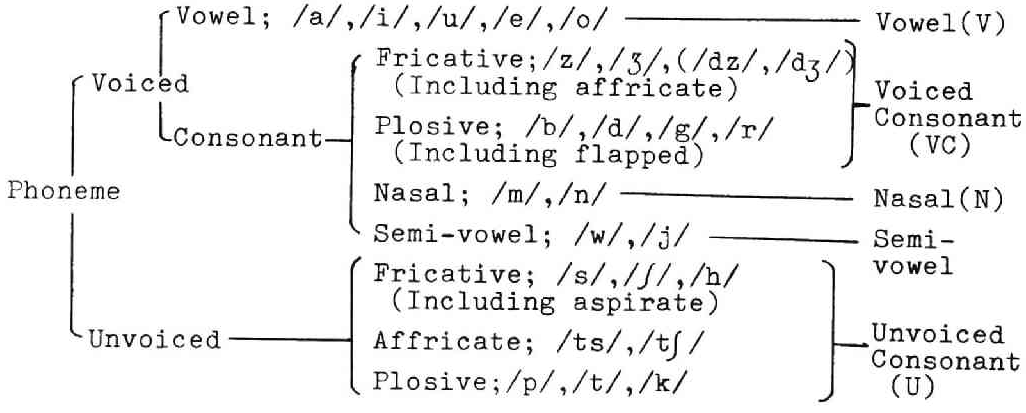
The segmentation signal which tells the detection of a new vowel segment is obtained by the logical combination of the above four signals.

3. Segmentation of Consonant and Vowel and Sampling Control

The method applied to the vowel segmentation can be applied to the segmentation between consonant section and vowel section. As both sections have the distinctively different characteristics with each other, their segmentation was performed by another way in segmentation circuit of consonant and vowel. The lower frequency components by vocal cord excitation and the higher frequency components by formant, hiss, etc. are detected, after the input speech sound is passed through a low pass filter and a high pass filter, respectively. By the logical combination of these signals consonant section is separated from vowel section.[25] (The circuits and the logics are explained in the later section with respect to the phoneme chassification.)

Sampling control circuit sends the sampling signal (A) to control the vowel analyzer and consonant analyzer in vowel section and in consonant section, respectively, and also sends the sampling signal (S) to the vowel segmentation circuit in vowel section.

Table 3.1 Classification of the Japanese phonemes. (The right column shows the classification in the phoneme classifier of Fig. 3.6.)



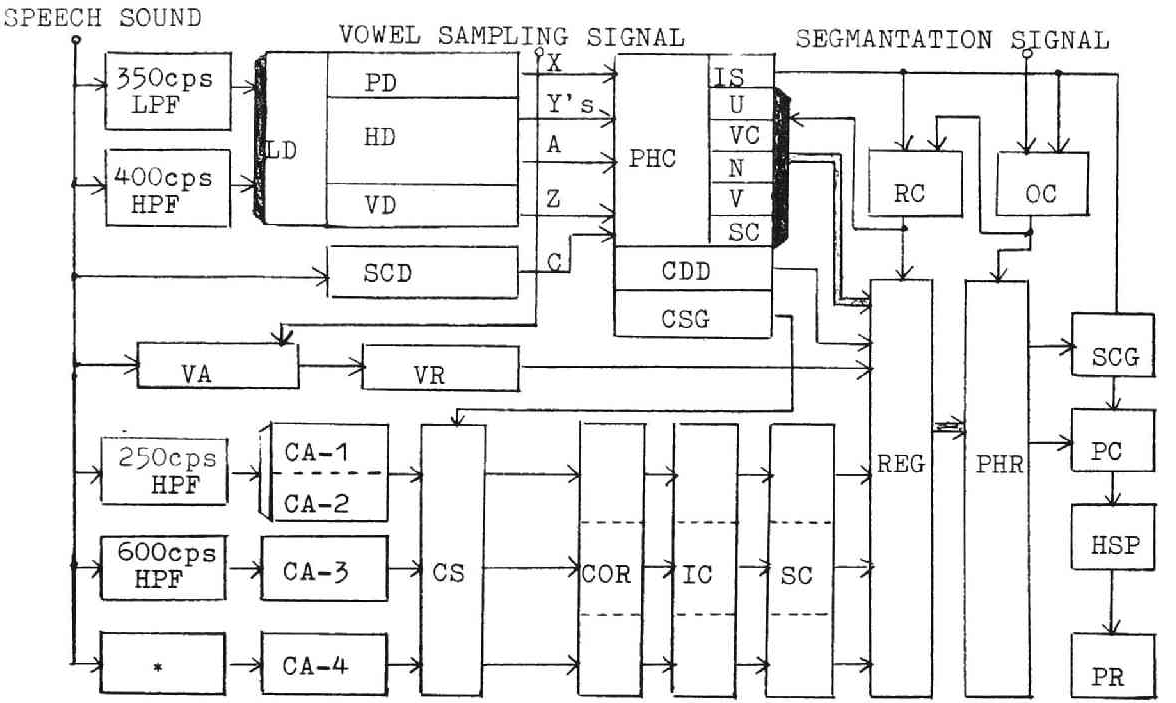
3·4 Recognition of Phonemes [25]

1. Feature Detection and Phoneme Classification

The speech sound is characterized in its pronouncing process by the manner and the place of articulation and, therefore, we designed the machine to treat the speech sound from these two aspects. That is, for the former we divide the segment of speech sound into several phoneme groups by the distinctive feature extractor and the phoneme classifier, and for the latter we discriminate the phonemes that belong to one phoneme group against each other by the analysis. The block diagram of this part is shown in Fig. 3·6.

The speech sound is classified into several groups as shown in Table 3·1 according to the distinctive differences characterized by the manner of articulation; they are vowel(V), unvoiced consonant (U), voiced consonant (VC) and nasal consonant (N). From the acoustic point of view the speech sound is grouped into sound segments of vowel, buzz sound, nasal murmur, burst and stationary noise sound. Each segment of the sound shows distinctive characteristics both in spectrum and in time domain. The vowel has the dominant formant components, while the level of consonant is in general weak and the spectrum is much complicated. In stop consonant the initial envelope shows abrupt building up. The voiced consonants other than nasal sounds have buzz sound which has dominant component in lower frequency.

As the level informations are perfectly omitted in the zero-crossing analysis, it is processed in the level detection circuits and therby the phoneme classification and the control of the operations are carried out. The outputs of a set of analyzing filters form the instantaneous representation of the speech sound. Although the increased number of filters can give the better representation, the complexity of the circuits will increase, too. In the recognition system described here, a rather simple method was adopted by taking the time variation of the levels into account.



SCG; special code generator, PC; puncher coder, HSP; high speed puncher, PR; printer, CA; consonant analyzer, CS; consonant sampling circuit, COR; channel OR gate, IC; integrating counter, SC; Schmitt circuit, *; filter for experimental use.

Fig. 3.6 Block diagram of recognition part.

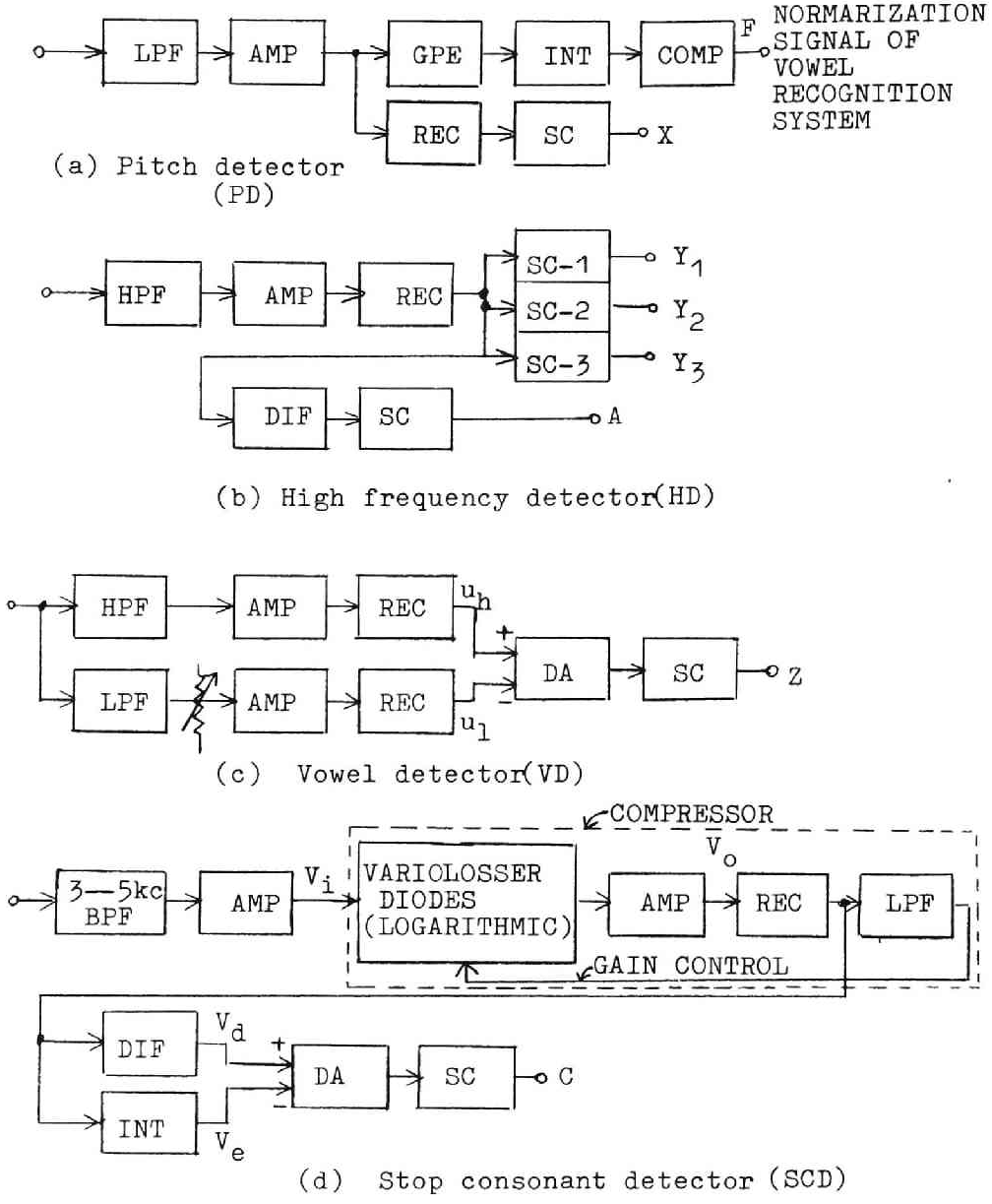
The components of speech sound are analyzed by low pass and high pass filters, from the outputs of which the speech sound is separated into several segments described above. Then from these segments, it is decided in which group of phoneme the speech sound belongs according to Table 3·1. At the same time the detection of stop consonant and the detection of the consonant duration are made.

The block diagram of the level detection circuits are shown in Fig. 3·7 (a)~(c). The speech sound is passed through low pass filter of 350 cps and high pass filter of 400 cps. The low pass filter detects the lower frequency components of voiced sound, while the high pass filter detects the high frequency components such as noise components, burst, formant components, etc.. The outputs of the circuits are the continuous logical variables.

Fig. 3·7 (a) is the pitch detector (PD). After amplification of low pass filter output, the envelope is detected by rectifier (REC), from which the binary variable X is obtained by Schmitt circuit. The output X=1, when the low frequency components exceed some preset level, corresponding to the voiced sounds such as vowel, buzz sound and nasal murmur. At the same time the pitch synchronous pulses are obtained from the Grüentz type pitch frequency extractor (GPE). The pitch frequency is then converted to corresponding analog voltage and averaged with large time constant to obtain the mean pitch frequency of the sound. By the voltage comparator (COMP) the frequency is divided into high and low ranges, by which the normalization of the vowel analysis circuit is controlled as described later.

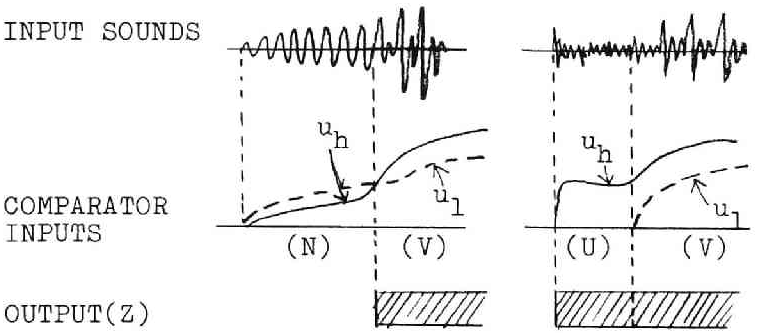
Fig. 3·7(b) is the high frequency detector. The level of the high frequency components are detected as the binary variable Y_1, Y_2 and Y_3. (They differ in the slicing levels of Schmitt circuits.)

The vowel detector of Fig. 3·7(c) uses the comparation of the high frequency components and the low frequency components, while the pitch detector and the high frequency detector use the absolute levels. The detected

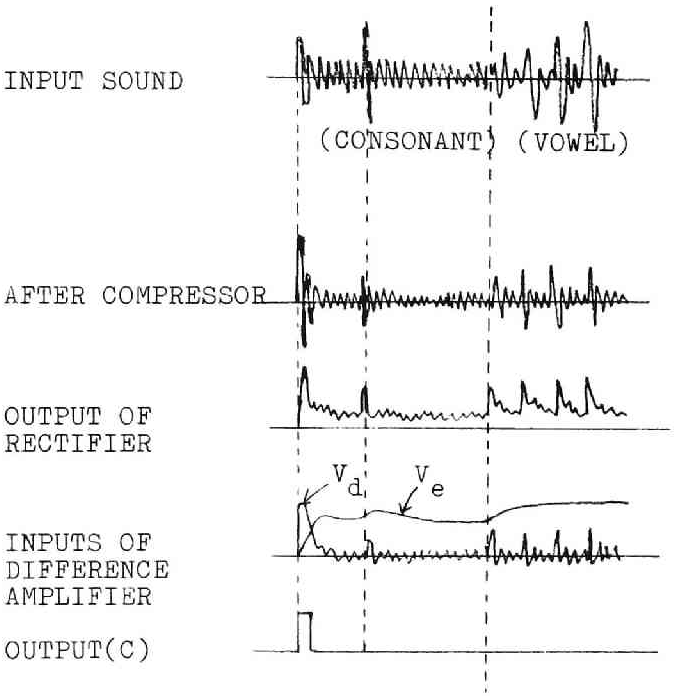


LPF; low pass filter(350cps), HPF; high pass filter(400cps), BPF; band pass filter, AMP; amplifier, REC; rectifier, SC; Schmitt circuit, INT; integrator, DIF; differentiator, COMP; comparator, GPE; Grüentz type pitch extractor, DA; difference amplifier.

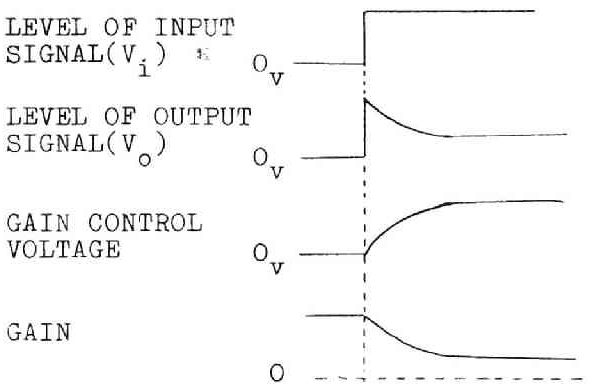
Fig. 3.7 Block diagram and operation of level detector and stop consonant detector. The inputs are the speech sounds and the outputs are logical variables. (See next page)



(e)Operation of vowel detector.



(g) Operation of stop consonant detector.



(f) Operation of syllable compressor.

Fig. 3.7 (Continued) The operations of the circuits.

levels u_h and u_ℓ of the outputs of HPF and LPF are compared in difference amplifier(DA). The binary output Z is 1 when $u_h > u_\ell$ and 0 otherwise. The level of the low pass filter output is appropriately adjusted in such way as Z=1 only for vowel sound. The circuit aims to distinguish the vowel sound and the nasal sound, in which the determination is difficult by absolute level processing. The operations of these circuits are shown in Fig. 3·7 (e)--(g).

The typical output responses of the variables X, Y's and Z are shown in Fig. 3·8 for various types of input monosyllables. From these variables, the speech sound is separated into several segments of vowel, unvoiced consonant (including noise part of voiced consonant), buzz sound and nasal murmur, by the logics XZ, $\overline{X}Y_1$, $X\overline{Y}_2$, and $XY_2\overline{Z}$, respectively. In vowel segment X and Y exist while in unvoiced consonant X is not detected. In syllable with voiced consonant, the first segment is buzz sound which is typically detected by $X\overline{Y}_2$ and the second segment is noise part which is detected by $\overline{X}Y_2$ in the same way as unvoiced consonant. The nasal murmur is detected by $XY_2\overline{Z}$. The phoneme classifier (PHC) is shown in Fig. 3·9, in which segment detection matrix, duration detection circuit and registers of phoneme groups, R-U, R-VC, etc., are provided. The matrix (a programmable diode matrix) performs the detection of segments stated above. If the duration of each segment continues longer than the interval set in the duration detection circuit, the segment is regarded as significant for the detection of a phoneme and the corresponding register is set. The outputs of registers are sent to main register for final phoneme recognition.

The states of registers are fed back to the matrix. The classification signal of vowel, V, is modified by these states as $XZ\overline{(U+VC+V)}$. That is, the variable V is inhibited, if one of the indication of consonants already exists, even when vowel segment is detected by XZ. Consequently the variable V will turn to "1", when the registers are reset after the phoneme recognition is executed while the vowel segment is being detected.

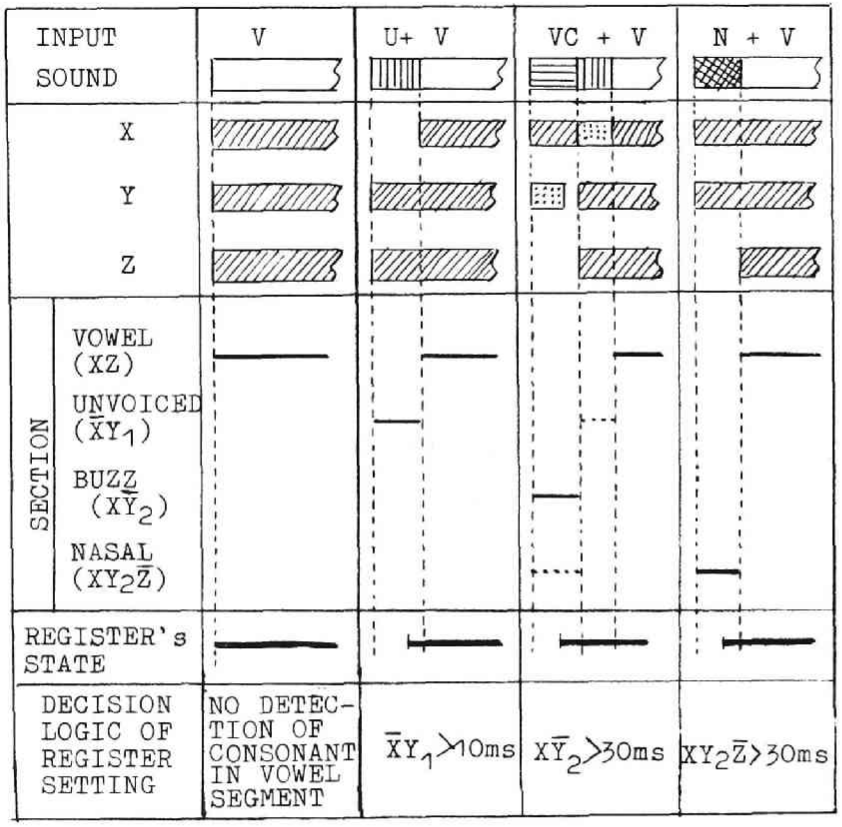


Fig. 3.8 Typical operation of phoneme classifier for several types of input sounds.

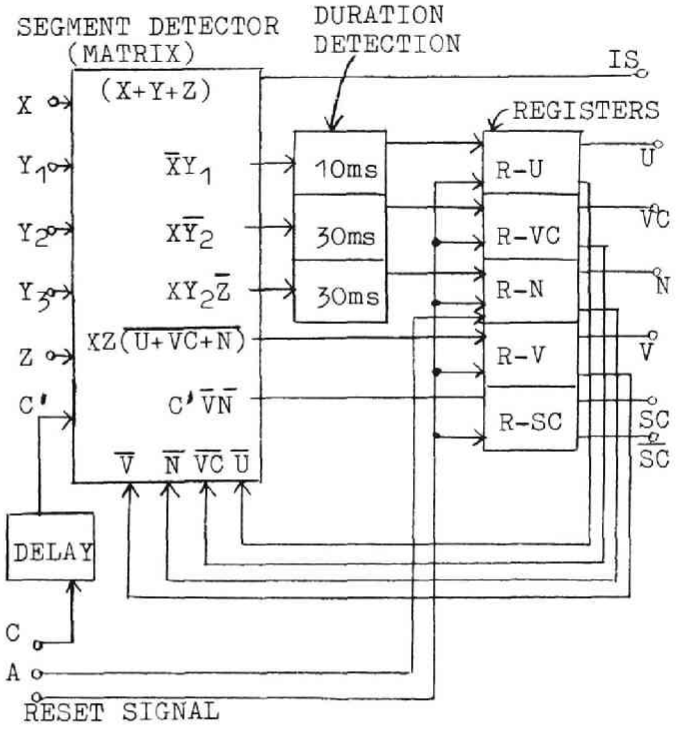


Fig. 3.9 Block diagram of phoneme classifier.

The signal of the speech sound period, IS, is obtained as X+Y+Z, which controls the over all operations of the whole system such as the reset of circuits, activation of the phoneme recognition circuit, detection of silent interval to put the space code in the output, etc.. The registers are reset at the initial of the signal lest the miss judgement by noise should interfere the operation for the comming sounds.

Buzz sound of voiced consonant often has a considerable higher frequency component comparable to the nasal murmur, especially in monosyllables. By the circuit stated above such voiced sound will be classified as nasal. The difference of both the voiced and the nasal in such case is that the level of the voiced consonant decreases before the burst, while in nasal the level increases gradually toward the following vowel sound. The decrease of level is detected in high frequency detector by differentiating the envelope and then detecting the negative slope.(Output A of Fig. 3·7(b)). When the impulse A is detected, the R-N register is reset and the R-VC register is set in consequence of it.

In parallel with the classification of phonemes, detections of the stop consonant and the duration of consonant segment are performed. Stop consonants have distinctive burst of envelope at the initial of the sound in the higher frequency region. The block diagram of stop consonant detector (SCD) is shown in Fig. 3·7(d). The 3—5 kc components of the speech sound are first amplified through syllable compressor and then, after the envelope is detected, it is differentiated by CR high pass network to detect the change of the envelope, on the one hand, and is integrated by CR low pass network to get the short-term averaged envelope on the other hand. The outputs of both networks, V_d and V_e, are compared by difference amplifier and the output is shaped by Schmitt circuit. The signal C is the pulse which appear when the burst is extracted. The R-SC register is set, if neither R-V nor R-N register is set at several tens milliseconds after the pulse C is generated.

The dynamic range of burst is considerably wide in conversational sound.

The syllable compressor can serve to reduce it. The steady state response is $Vo = k\sqrt{Vi}$ (k;constant, Vi;input level, Vo;output level), using logarithmic diodes. Another aspect of the compressor is the emphasis of the initial onset of the sound. Because of the low pass filter in feedback loop the onset of signal will be emphasized, whereas the spikes in the continuous back ground signal are weakened.

The detection errors of the stop consonant detector described here for the unvoiced consonants (54 stop consonants and 20 noise consonants of male and female sound appered in connected speech) were 2 against the 10 dB shift of the input level by the attenuator and 4 against the 20 dB shift.

The duration of consonant is measured by counting 1 kc pulses with integrating counter. When it exceeds a preset value (e.g., 30ms) one of the registers is set, which is used for the discrimination between /k/ and /t/, and /d/ and /g/.

The consonant segment signal $\overline{XY}_3$ is also used as sampling signal of consonant, by which noise part of unvoiced and voiced consonant is sampled. Another sampling signal is derived from it in such way as it does not exceed the predetermined duration (e.g., 40ms).

The principle of the phoneme classifier adopted is rather simple. For the elaborate operation it will be desirable that the number of filters is increased and the amplitudes must be processed with relative relation in dB scale by procedures such as spectrum matching in which time variation must be handled as well. At present stage the level of the input speech sound must be adjusted as much as possible, especially in connected speech. For the separation between vowel sound and the lax /p/ sound having no aspiration, detailed spectral structure and the periodicity of the signal must be utilized (see chapter 4, PART I). The method explained in section 4·2, §5 of PART I will be available for the separation between the vowel and the nasal murmur.

2. Recognition of Vowels

The vowels are the most important phonemes in speech sound. In Japanese we have five vowels; /a/, /i/, /u/, /e/ and /o/. They occupy the greater part of the speech sound, having the dominant power. The vowel is generated by exciting the vocal tract with the periodic buzz source of constant volume velocity and radiated from the mouth opening. The spectrum has several complex zeros by source wave form. The dominant features of vowel phonemes are, however, the peaks or resonances of spectrum, called formants. The frequencies of the formants depend not only on the vowel phoneme but also on difference of speaker, especially on the difference of sex, which causes some difficulties in the processing of vowels. In general the formant frequencies of female voice are higher than those of male voice. In recognition such deviation of formant frequencies must be normalized. From the spectrographic data of vowels, it is known that the formant frequencies of the first formant, the second formant and the third formant (F_1, F_2 and F_3, respectively) changes proportionally when the pitch frequency changes.[26] It is theoretically deduced by Kondo that the relative values of f_1, f_2 and f_3 are the parameters for vowel recognition.[27]

It is needed for the recognition of vowel to perform the extraction of formant frequencies, the normalization and the partition of domain into vowel phonemes. The most powerful scheme of formant extraction is the spectrum matching using the analysis by sythesis method.[28] The method is, however, too complicated to realize it in real time system using a special circuits.

The vowel recognition circuit of this speech recognition system is based on the zero-crossing analysis, which was explained in chapter 2. The vowel recognition circuit successively works during the vowel section, controlled by the sampling singnal. There are prepared two zero-crossing analysis circuits for the first formant and for the second formant, respectively, each of which detects the formant frequency and then based on the frequencies the vowel recognition is executed. The deviation of formant frequencies by the male and the female voice is normalized by selecting the circuit condition with

the pitch frequency.

In zero-crossing analysis of Fig. 2·5 the channel classification characteristics can be changed proportionally by changing the clock pulse frequency f_c. Therefore, when the formants of some distribution are proportionally changed from the formants of standard distribution, the distribution can be normalized to the standard by changing f_c by the corresponding value. Some problems arise as for the filter used in the experiment to separate the formant components. For normalization the cut off frequency of the filter must also be changed. This requirement can be satisfied in the first formant circuit by choosing the cut off frequency rather high. As in the second formant the band pass filter is used and its cut off frequencies are critical, the normalization of filter characteristics is needed. In this experiment the pass band of band pass filter of the second formant is effectively shifted by shifting the frequency of the signal before the conversion to zero-crossing wave. The shifting of signal is performed by the SSB modulation and the demodulation. The shifted frequency f_s is the difference of frequencies of both carriers.

The other problem is the selection of the parameter to control the normalization operation. One possible way is to use the third formant frequency f_3, in which method the parameters of recognition are f_1/f_3 and f_2/f_3, but the exact extraction of f_3 is difficult. Another way is to use the pitch frequency f_p, because formant frequencies and f_p are nearly proportional. In the experiment the circuit was switched according as the pith frequency is high (female voice) or low (male voice).[30]

The perfect separation of formant components is impossible by the fixed filter. In the first formant region, the second formant of /u/, /o/ will interfere the exact extraction of the first formant. In /a/, the first and the second formants (F_1 and F_2) are extracted as one formant. In the second formant region, F_1 and F_3 will intervene, especially for /u/ in which the

second formant level is often weak.

The zero-crossing distribrtion of the first formant region $W_i^{(1)}$ and the second formant region $W_i^{(2)}$ are shown in Fig. 3·10 and Fig. 3·11, respectively. The circuits used are the same as shown in Fig. 3·2. The speech samples are Japanese monosyllables. The sampling for analysis was made in the middle part of vowel section. The zero-crossing distribution $W_i^{(1)}$ of 8 channels were obtained after the signal was passed through low pass filter of 0～1500 cps. As for the second formant, the signal was shifted by f_s cps before passing it through the band pass filter. The channel characteristics are shown in Table 3·2. In most cases the first formant distribution of /u/ and /o/ is rather complicated and the formant frequencies of female voice are higher than in the male voice. The second formant distribution of female voice for $f_s = 0$ is quite useless and $f_s = 300$ cps or $f_s = 400$ cps will give better representation.

Table 3·2 Frequency characteristics of the channels of zero-crossing analysis in cps.

(a) FIRST FORMANT

CHANNEL NUMBER	1	2	3	4	5	6	7	8
FREQUENCY CHARACTERISTICS	165 | 340	340 | 405	405 | 470	470 | 545	545 | 690	690 | 800	800 | 940	940 | 1450

(b) SECOND FORMANT (without frequency shift)

CHANNEL NUMBER	1	2	3	4	5	6	7	8	9
FREQUENCY CHARACTERISTICS	700 | 800	800 | 940	940 | 1140	1140 | 1270	1270 | 1450	1450 | 1650	1650 | 2000	2000 | 2500	2500 | 3600

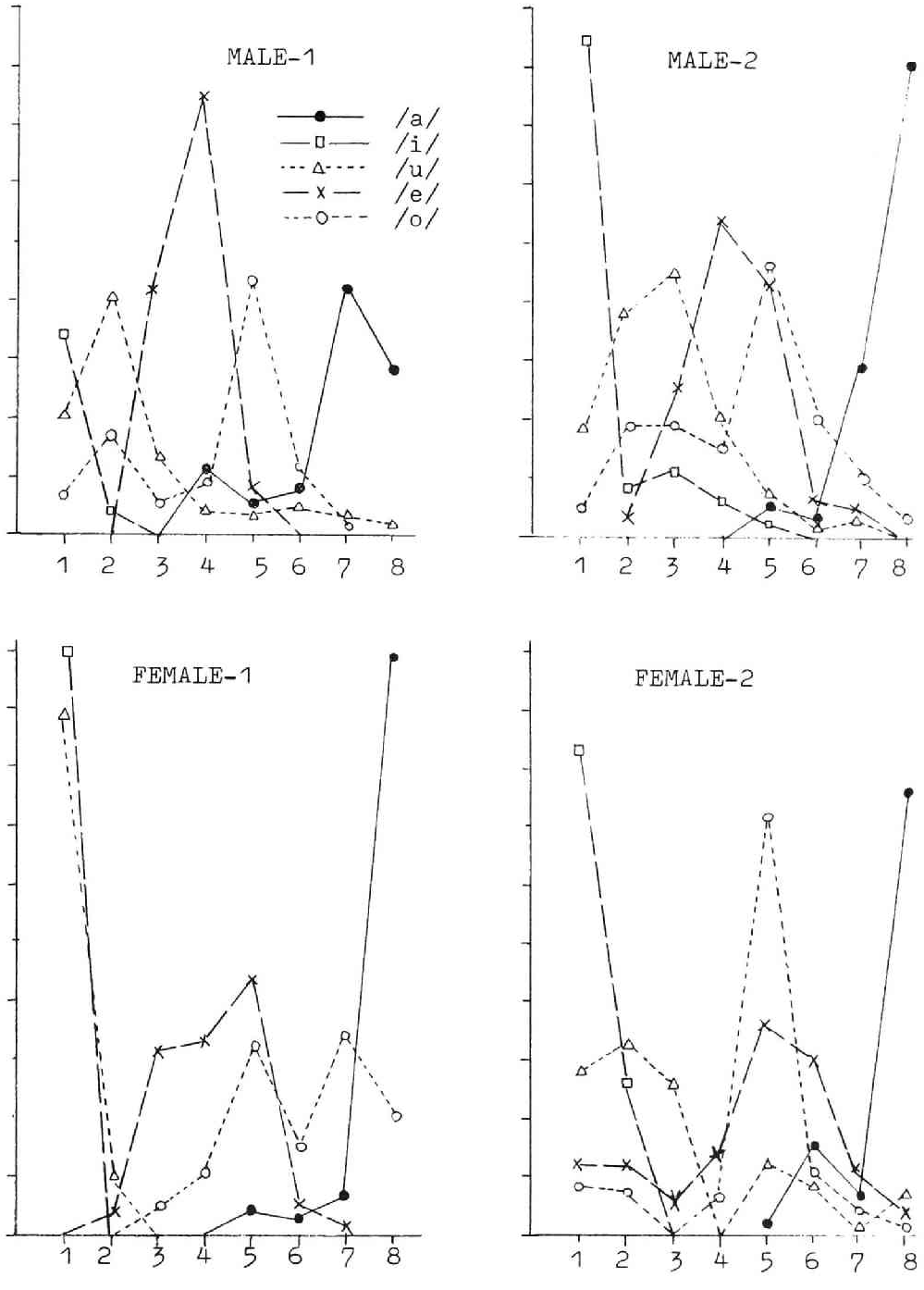


Fig. 3.10 First formant zero-crossing distribution of Japanese vowels.
Abscissa; channel number(As for the characteristics see Table 3.2)
Ordinate; relative zero-crossing distribution.

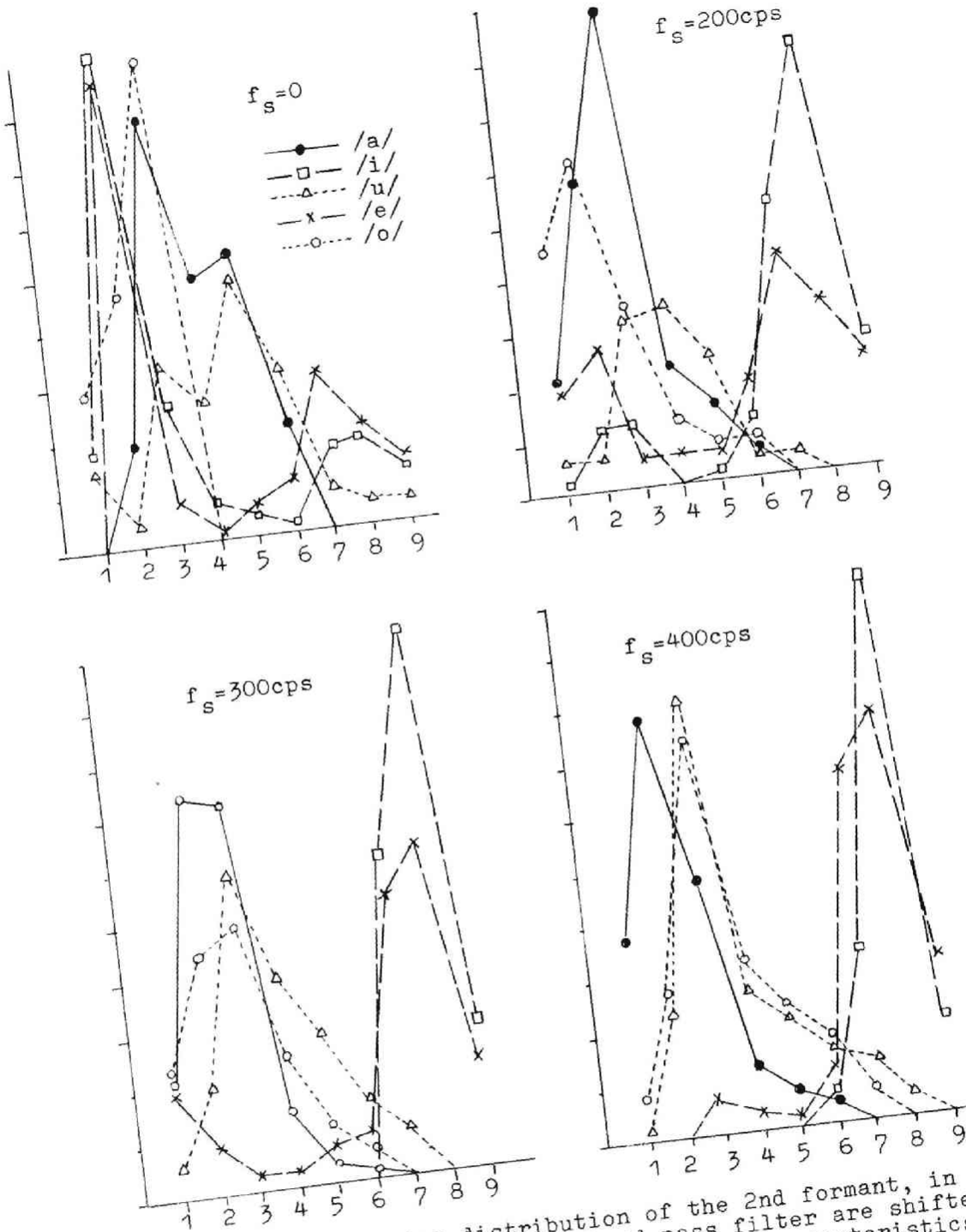


Fig. 3.11 Zero-crossing distribution of the 2nd formant, in which the characteristics of the band pass filter are shifted fs cps. Abscissa; channel number (As for the characteristics see Table 3.2.) Ordinate; relative zero-crossing distribution.

Fig. 3·12 shows the block diagram of the vowel recognition system used in speech recognition system, in which three zero-crossing circuits (the two for the first formant region and the one for the second formant region) and vowel recognizer are prepared. The two circuits differ in its cutoff frequencies of low pass filters, i.e., 850 cps for No.1 circuit and 1500 cps for No.2 circuit. The normalization is carried out by switching the clock frequencies of zero-crossing measurement circuits, f_{c1} and f_{c2}, and the shift frequency f_s. The sampling is repeated every 20 ms. The operations of zero-crossing analysis circuit, recognition matrix and vowel segmentation circuit are synchronous.

The vowel recognition in F_1-F_2 domain is possible by detecting the peak channels from the distribution of F_1 and F_2. The method is, however, inappropriate, because the form of distribution is not so simple and also the domain partition best for the recognition differs for each vowel. For these reasons the logical variables particular for each pair of phonemes were used in F_1 region.

In Fig. 3·12, X_1, X_2, X_3, X_4 and Y_1, Y_2, Y_3 are the logical variables for the first formant region and for the second formant region, respectively. To obtain these variables, the zero-crossing distributions are merged by OR gates. After that the merged distribution W_1, W_1', etc. are converted to voltage by integrating counters. The variables are formed by comparators as follows:

X_1(/e/ /i/) 1 ; $W_1(850-365) \geq \alpha_1 W_1'$ (365 -- 175)

X_2(/e/,/o/ - /u/) 1; $W_2(850 - 490) \geq \alpha_2 W_2'$(490 -- 175)

X_3(/o/ - /u/) 1 ; $W_3(850 - 430) \geq \alpha_3 W_3'$(430 -- 175)

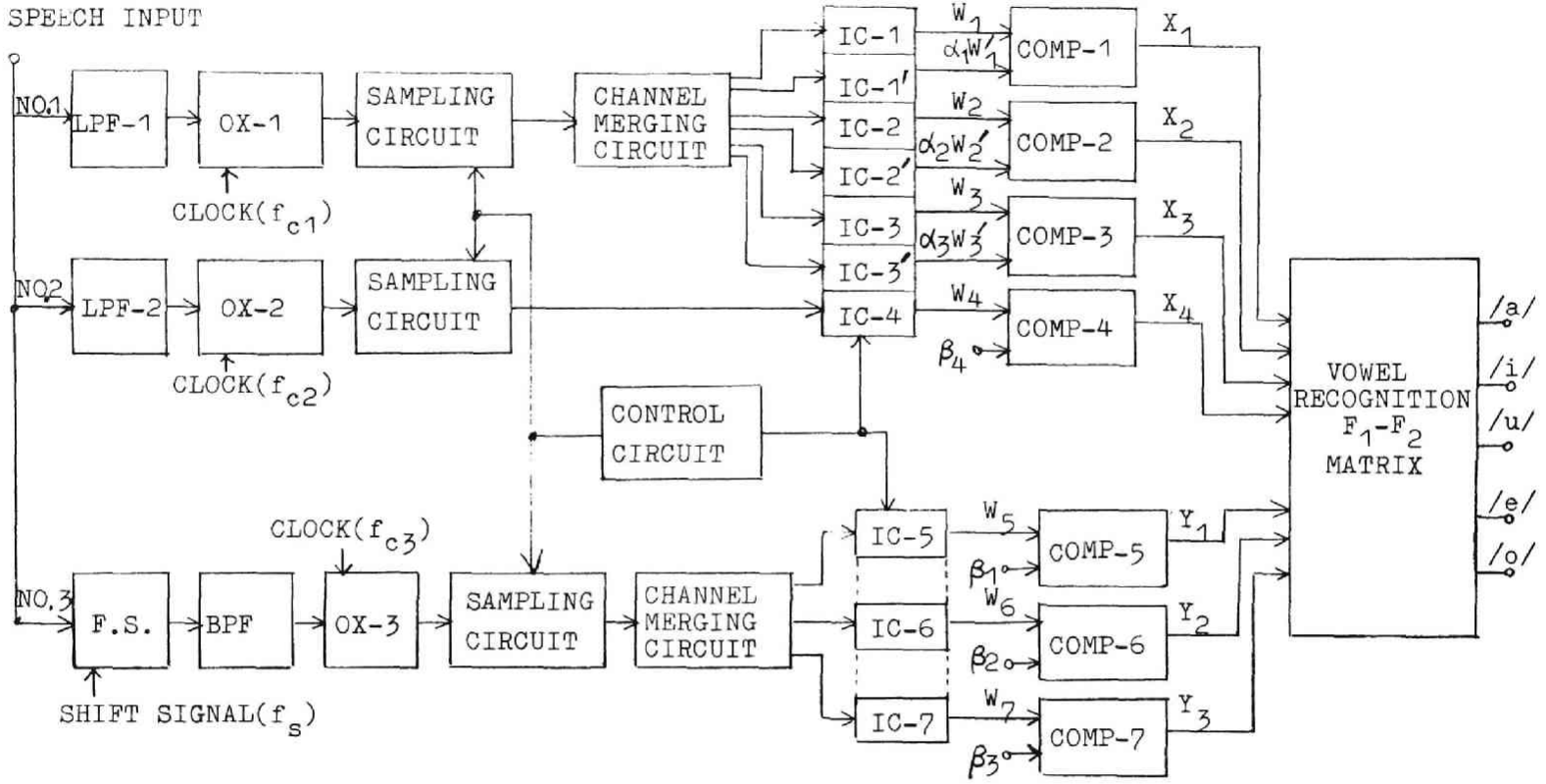
X_4(/a/ - /o/,/u/) 1; $W_4(1450 - 750) \geq \beta_4$

$Y_1 = 1$; $W_5(3600 - 1700) \geq \beta_1$

Y_2 1 ; $W_6(3600 - 1460) \geq \beta_2$

Y_3 1 ; $W_7(1040 - 720) \geq \beta_3$

(Each variable is zero, if the respective condition is not satisfied.)



LPF-1; 800cps, LPF-2; 1500cps, BPF; 800~2500cps, F.S.;frequency shift circuit, OX; zero—crossing analyzer(c.f. Fig. 2.4), IC; integrating counters, COMP; comparator.

Fig. 3.12 Block diagram of the vowel recognition system.The operations of sampling circuit, integrating counters etc. are performed in mode A and B in the same way as explained in Fig. 2.4,controlled by the control circuit.

in which;

W(850 365) means the value of distribution of zero-crossing wave in the range of 850 cps$\sim$365 cps, and so on.

α_1, α_2 and α_3 are weighting constants and $\beta_1 \sim \beta_4$ are threshold constants experimentally decided for each phoneme. X_1(/e/-/i/) means that the variable X_1 is used for the discrimination between /i/ and /e/ and that X_1= 1 corresponds to /e/ and X_1 = 0 to /i/, and so on.

As seen from Fig. 3·12 X_1, X_2 and X_3 are obtained from circuit No.1 and X_4 from No.2. For simplicity some distributions are compared with constant levels β's.

New logical variable Z's are derived from Y_1, Y_2 and Y_3.

$Z_1 = Y_1$, $Z_2 \quad \overline{Y}_1 Y_2$

$Z_3 = \overline{Y}_2 \overline{Y}_3$, $Z_4 = \overline{Y}_2 Y_3$,

in which it is designed so as to hold the relation $Y_1 C Y_2$.

Using X's and Z's, the decision logic is composed as shown in Table 3·3. The unique selection of a mesh is accomplished by the variable Z's, although $X_1, \cdots\cdots, X_4$ are not disjoint. Each mesh is assigned for the vowel phoneme written in it. The particular variables Xi of the F_1 region are combined with particular variable of the F_2 region. For the instance of Z_4=1, if X_4 =1 then the vowel sound is recognized as /a/. if X_4=0, X_3=1 as /o/ and if X_4= 0, X_3=0 as /u/.

The results of recognition of vowel sound in monosyllables are shown in Table 3·4 (a) for the male voice and in Table 3.4.(b) for the female voice. The sampling was made successively with the interval of 20 ms during the vowel sound except the initial 30 ms. For each sampling, recognition among five vowels were made, the results of which were accumulated by the counters, each prepared for a particular vowel phoneme, all over the one vowel sound.

The recognition was judged to be correct when (1) the number of correct recognition is greater than 2, (2) it exceeds the number of wrong recognitions by more than 2, and (3) the number of wrong recognitions does not exceed 5.

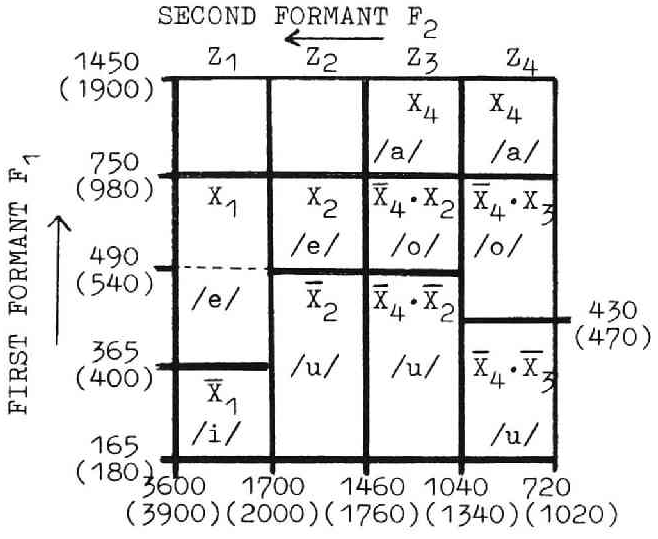


Table 3.3 Vowel recognition matrix in F_1-F_2 domain. Frequency characteristics are shown in cps for male voice (low pitch selection), while for female voice (high pitch selection) they are shown in bracket.

RECOGNIZED

ARTICULATED	a	i	u	e	o
a					25
i			4		
u				3	10
e	1		3		
o	8		4		

229 SOUNDS FOR EACH PHONEME
1145 IN TOTAL
11 SPEAKERS

(a) MALE VOICE

RECOGNIZED

ARTICULATED	a	i	u	e	o
a					11
i			1	18	
u					20
e			9		
o	2		19		

168 SOUNDS FOR EACH PHONEME
840 IN TOTAL
9 SPEAKERS

(b) FEMALE VOICE

Table 3.4 Confusion matrices of vowel recognition.

For male voice, the average score is more than 94% for 11 speakers. The confusion between /a/ and /o/ is caused by a particular speaker. The confusion between /u/ and /o/ occurs in F_1 region, in which the detection of F_1 is troubled by the intervention of the F_2 components.

For female voice the average score is about 90% for 9 speakers. The confusions between /u/ and /o/, and /i/ and /e/ are caused by a particular speaker. The confusion from /e/ to /u/ occurs in F_2 region.

Some of the mis-recognized sound were found to have vague perception. Therefore the score of recognition may be improved by a little training of speaker so as to make clear articulation.

3. Recognition of Consonants

The analysis of consonant is carried out by zero-crossing analysis. In parallel with zero-crossing analysis, the several features are extracted in phoneme classifier (PHC). By combining these results the consonant recognition is performed in phoneme recognition circuit (PHR) in combination with vowel recognition.

The acoustic properties of consonants are so different according to the phoneme group because of the difference of the manner of articulation. Therefore, the description of consonant is not so simple as the vowel is perfectly described by formant structure. As well as the power spectrum, statistic characteristics of noise, duration of consonant, envelope information, etc. must be considered.

The consonants suffer much interaction from the adjacent vowels which is expressed as the transition from consonant to vowel or vice versa. The informations of consonant are, therefore, in the stationary noise part like hiss, burst, etc. and in the transition part.

The features of unvoiced consonant are mainly included in the noise part (the transition appears as aspiration), while the transition part of

voiced consonant is important feature of recognition, whereas the noise is weaker than in unvoiced consonant. The transition is effective in stop consonants (/b/, /d/, etc.) in which the movement of speech organ toward the vowel is remarkable. In nasal sound the burst by the opening of the mouth closure continues to vowel sound. The unvoiced and the voiced consonant have almost one to one corespondence such as /p/ and /b/, /t/ and /d/, etc., although the noise or burst of voiced consonant are not remarkable as in unvoiced consonant. The buzz sound is usually observed before the start of utterance of noise or burst sound (Here only the noise part except the buzz sound is called consonant part which is the objective of the analysis.) The principle of recognition of the unvoiced and the voiced is the same, though in voiced consonant the utilization of transition is useful.

As the consonants are grouped into the unvoiced, the voiced and the nasal consonant, the analysis and recognition among the phonemes that belong in one phoneme group are made in the separate circuit.

The block diagram of consonant analyzing system is shown in Fig. 3·6. The consonant analyzer (CA) has the same constitution as the vowel analyzer (VA), as have been explained in chapter 2. In this case, however, the input filter is adjusted for each analysis and the channel classification characteristics are different. The consonant segment is sampled once for the full duration or for a limited duration. The zero-crossing distribution obtained in consonant snalyzer is classified into a number of channels (e.g., 14 or 16). The channels are then merged by OR gates to obtain the zero-crossing distribution suitable for the recognition of a pair of phonemes of interest. (The merging is also reasonable because the zero-crossing distributions of consonants are in general spread or not so compact.) The merged distributions are then expressed as voltages using integrating counters

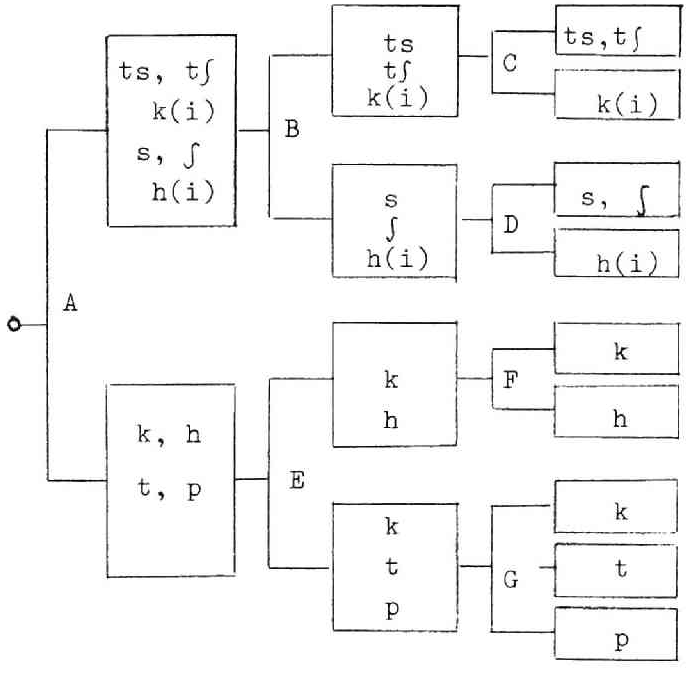


Fig. 3.13 Constitution of the decision logic of unvoiced consonant. A,B,···,G are the branch identifiers.

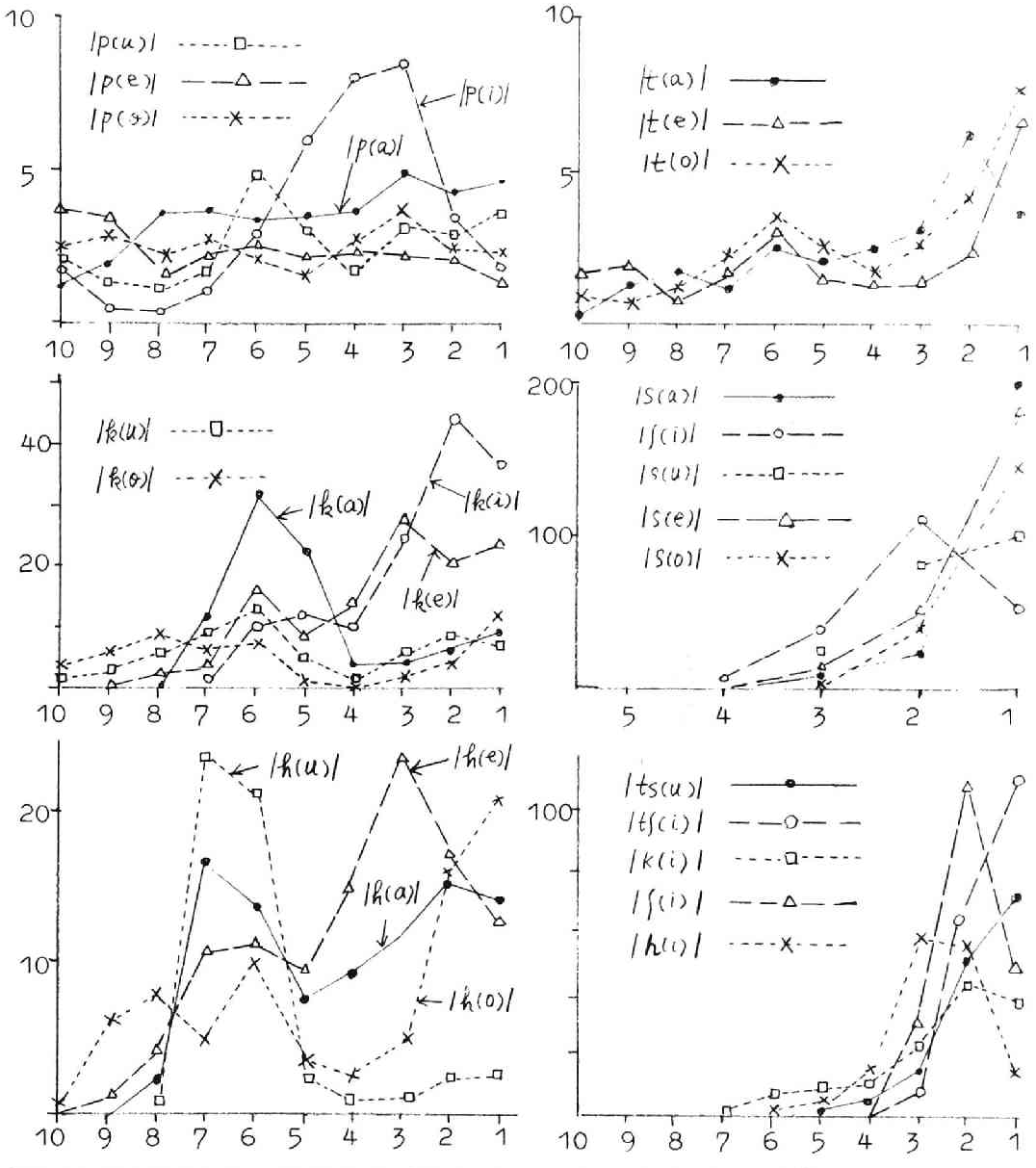
(IC), the conversion rate from pulse number from the zero-crossing analyzer to the voltage can be adjusted for each purpose. Unlike the vowel recognition the voltages are converted to binary signal by Schmitt circuits, the threshold levels of which are decided from the data. The binary signals are used as the logical variables for consonant recognition in phoneme recognition circuit.

The threshold level logic is more suitable for consonant than the peak detection because the size of the distribution is also the parameter of recognition. However, in some cases the comparison of two outputs of the merged zero-crossing distributions are performed by differential integrating counter (the integrating counter with two inputs, for one of which the counter operates as adder and for the other of which it operates as subtractor.)

The constitution of the recognition logic for unvoiced consonant is shown in Fig. 3·13, in which the features such as the duration of consonant, stop consonant detection, etc. are utilized as well as the zero-crossing analysis. The logics form tree system, but a paricular feature (variable) is used for the decision logic of each branch indicated by A, B, etc.. For instance there are two variables on stop consonant detection, B and F. The decision logic is not perfect tree system; /k(i)/ and /h(i)/ are branched into both groups by the first stage logic A and they are treated in each branch separately, the results of which are gathered after the recognition.

The zero-crossing distributions of unvoiced consonants are shown in Fig. 3·14 for /k/, /t/, /p/, /s/, /ʃ/, /h/, /ts/ and /tʃ/. The data were obtaind for monosyllabic sounds of one speaker by passing the high pass filter of 250 cps cutoff frequency. The distribution is affected by how the articulation is made rather than by the difference of speakers.

In /p/ and /t/, distributions show no characteristic shapes such as peaks, being distributed in wide range. Though the higher components of /t/ is a little stronger than of /p/, the difference is not clear. This is because /p/ and /t/ are characterized by short transient impulse signal, which

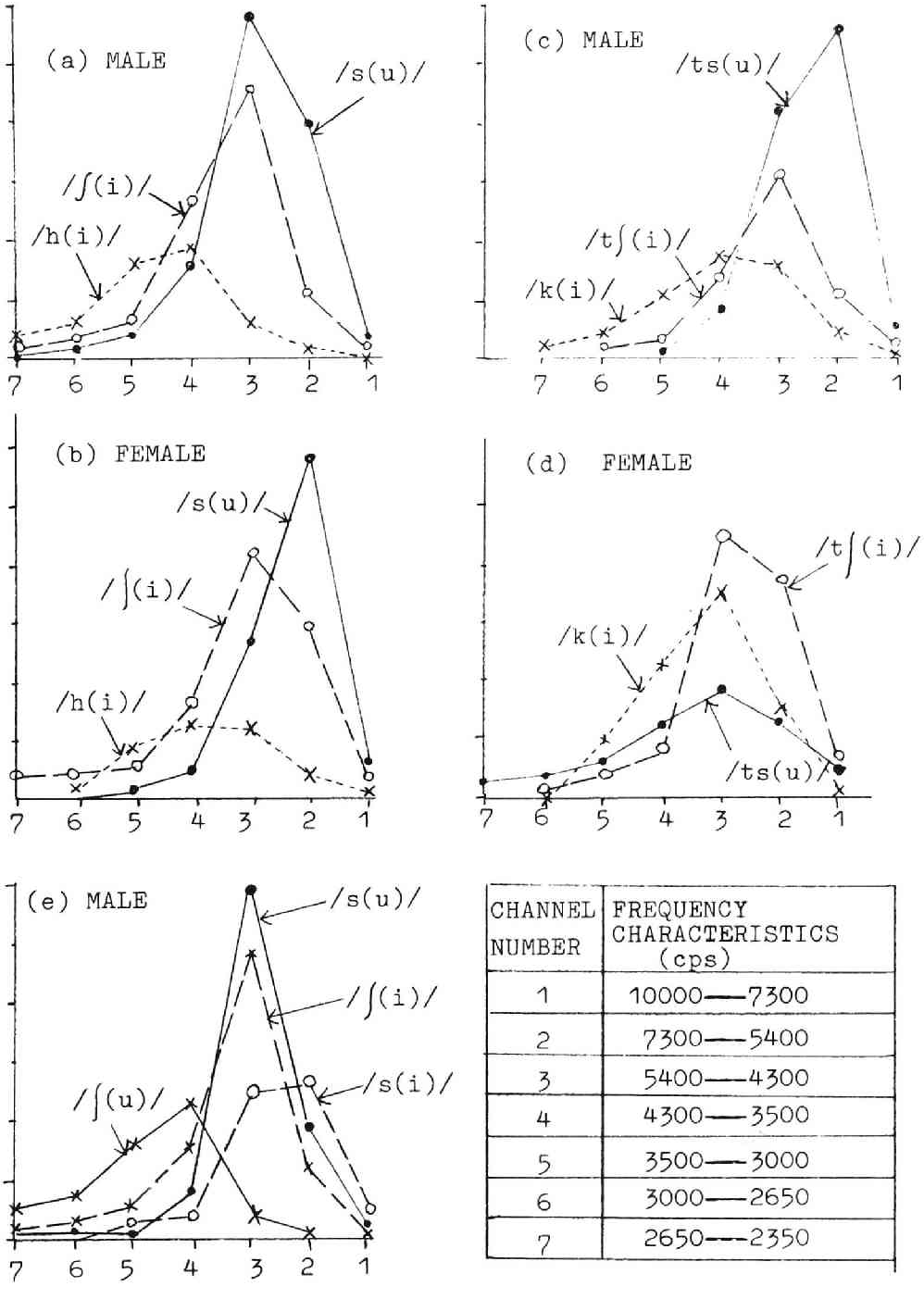


CHANNEL NUMBER	FREQUENCY CHARACTERISTICS (cps)				
1	9600---5000	3	3340---2500	7	1250---920
2	5000---3340	4	2500---2000	8	920---680
		5	2000---1670	9	680---420
		6	1670---1250	10	420---200

Fig. 3.14 Distribution of zero-crossing frequency of unvoiced consonant(for one speaker). The sampling was made for less than 40ms in the consonant segment.
Abscissa; Channel number, Ordinate; Frequency of zero-crossing.

hardly contributes to zero-crossing distribution. The peaks of /k/ and /h/ due to the aspiration corresponding to the second or the higher formant are affected by the following vowel. The distributions of /k/ and /h/ followed by the same vowel are similar, although the formant is conspicuous in /h/. The distributions of fricative and affricate consonants including /h(i)/ and /k(i)/ are concentrated in higher frequency region. The peak of such fricative, affricate and aspirate consonants characterized by a dominant formant correspond to the major formant frequency and also to the mean zero-crossing interval,[20] and the weak formants and the antiformants can not be presented. The peak of fricative consonant is situated in the frequency region above 3·34 kc, the position of which depends on the following vowel; that is, in /ʃ(i)/ and /s(u)/it is lower than the others.

At branch A of Fig. 3·13 the consonants are separated into /s/, /ʃ/, /tʃ/ and /ts/ (hiss group) and /h/, /k/, /t/ and /p/ according as the channel outputs of the zero-crossing distribution appeared in the range above 3·3 kc exceed a threshold level or not(hiss detection). /k/ and /h/ sound is usually classified into non-hiss group but /k(i)/ (/k/ followed by /i/) and /h(i)/ often confused as hiss sound. Therefore the perfect decision of these is not intended at this step. At step B the output of stop consonant detector (SC-1) branches /ts/, /tʃ/ and /k(i)/ (stop consonant group) and /s/, /ʃ/ and /h(i)/ as described previously. The separations between /s/ and /ʃ/, and /ʃ/ and /h(i)/ are rather difficult problems (step D). The zero-crossing distributions of /s(u)/, /ʃ(i)/ and /h(i)/ are shown in Fig. 3·15 (a) and (b) for male and female voices. The distinction among these phonemes are possible for one speaker, but not for several speakers. The situation is the same for group /ts(u)/, /tʃ(i)/ and /k(i)/ as shown in Fig. 3·15 (c) and (d). (In (d), the major formant of /ts(u)/ exists in higher frequency than the cutoff frequency of the filter, 7kc, by which the distribution is deformed considerably.) In Fig. 3·15 (e) the results of /s/ and /ʃ/ followed by vowels /u/ and /i/ are presented, although /s(i)/ does not exist in Japanese. The distinction



CHANNEL NUMBER	FREQUENCY CHARACTERISTICS (cps)
1	10000—7300
2	7300—5400
3	5400—4300
4	4300—3500
5	3500—3000
6	3000—2650
7	2650—2350

Fig 3.15 Zero-crossing distribution of fricative and affricate consonants. The signal was passed through low pass filter with cutoff frequency of 7kc and sampled in full segment of the consonant.

between /ʃ(u)/ and /s(u)/ will be established but /s(u)/ and /ʃ(i)/ show quite the same distributions.

The distinction between /ʃ(i)/ and /h(i)/ may be possible from the values of distribution of channel No.2 and No.3 or from the difference of the magnitude of distribution which is due to the difference of the consonant durations. The distinction between /tʃ(i)/ and /k(i)/ is difficult from these data, which was tried by the detection of the envelope (step C) as follows.

The block diagram of the discrimination circuit is shown in Fig. 3·16 (a) and the operation in (b). The principle is to compare the energy at the initial part of the sound (burst) and the energy during the stationary noise part (aspiration). The ratio of the former to the latter will be greater in /k(i)/ than in /ts(u)/ or /tʃ(i)/. in which the burst is sharp and the fricative noise level is high. The circuit, therefore, samples the energy of the initial part and holds it by sample and hold circuit (SH), and the noise energy is integrated by integrator (INT) and then both voltages, V_H and V_I, are compared. The circuit starts its operation by the detection of the furst from the stop consonant detector and judges the voltage of comparator after the integrating interval set in monostable multivibrator (MM-3) or when consonant segment has been finished. The circuit is designed so as to generate the output signal D for input sound /k(i)/. The signal D is led to the phoneme recognition circuit (PHR) of Fig. 3·6. The average sore of the circuit for /ts(u)/, /tʃ(i)/ and /k(i)/ appeared in monosyllable spoken by four speakers is about 90%. (As for the zero-crossing analysis of fricative consonants by means of single tuned filter, refer to section 2·4)

On the other hand the group of /h/, /k/, /t/ and /p/ is separated into /h/, /k/ group and /k/, /t/ and /p/ group by consonant duration (Step E).[29] As the duration of /k/ is somewhat shorter than /h/, /k/ may be classified into both groups. The distinction between /h/ and /k/ is performed by stop consonant detection (SC-II) at step F, the circuit of which has different circuit constants from those of SC-I. At step G the group /k/, /t/ and /p/

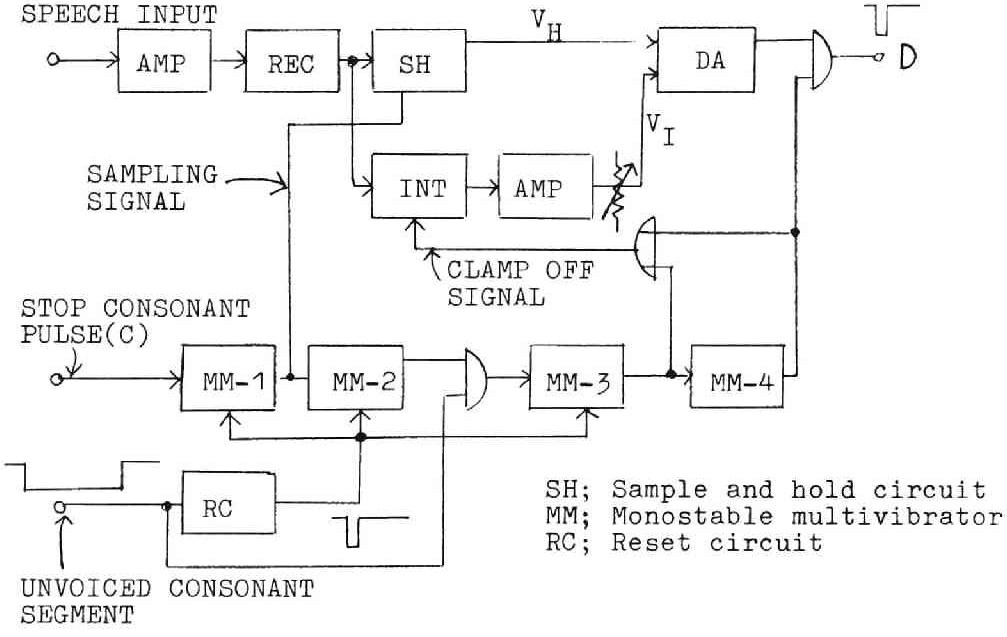


Fig. 3.16(a) Block diagram of discriminating circuit of /ts(u)/, /t∫(i)/ and /k(i)/.

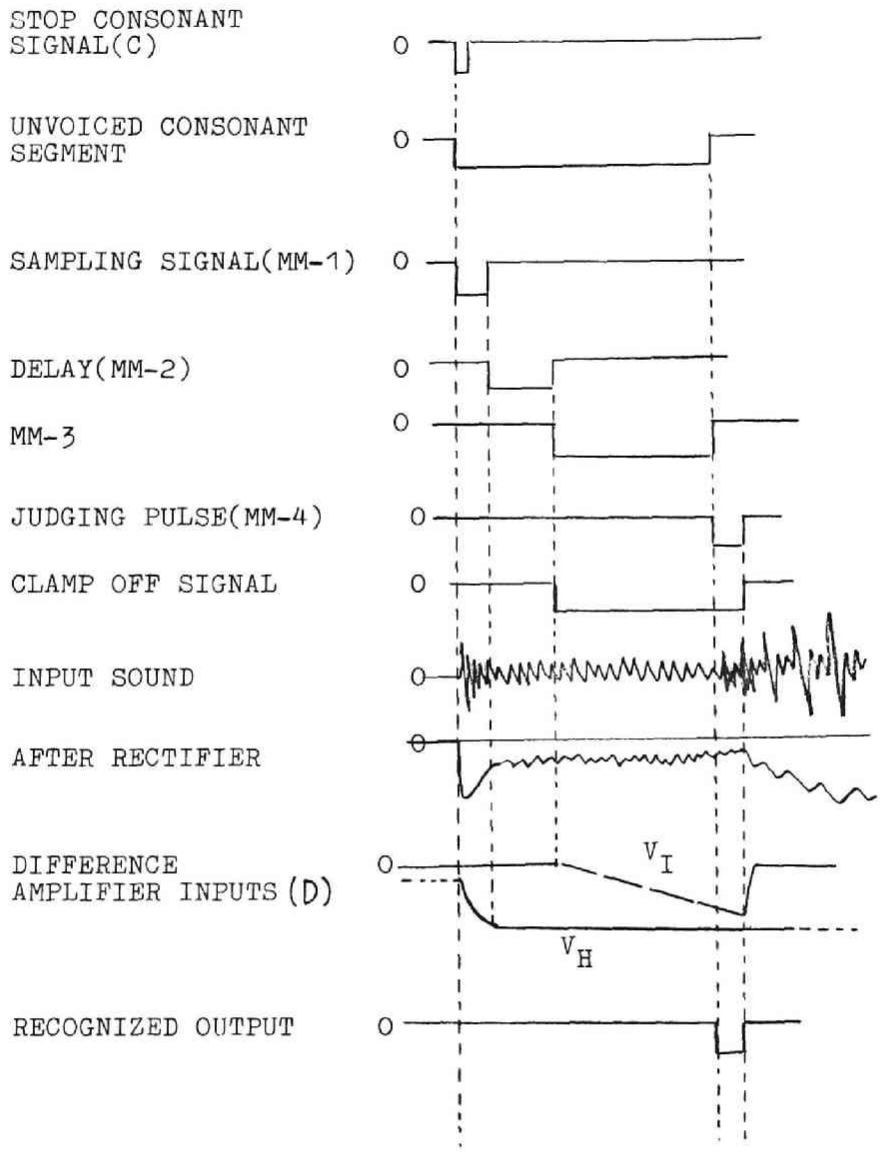


Fig. 3.16(b) Operation of discriminating circuit of /ts(u)/, /t∫(i)/ and /k(i)/ for the input sound /k(i)/.

is separated using the zero-crossing distribution. The distribution suffers the influence from the following vowel. The discrimination is, therfore, performed among the consonant class followed by the same vowel; that is, among /k(a)/, /t(a)/ and /p(a)/, etc.. For each class, there are two logical variables derived from merged zero-crossing distribution. The distinction between /t/ and /p/ is left unsolved.

The average score of unvoiced consonant recognition is about 70%, though the detection score of each feature used in each step of recognition tree is better than that. The sore largely depends on the manner in the utterance of the sound. It must be clear as possible. The confusions occur between /p/ and /t/, /k(i)/ and /tʃ(i)/, /h(i)/ and /ʃ(i)/ and also between /p/, /t/ and vowel.

The constitution of the voiced consonant recognition resembles to that of the unvoiced consonant with some modifications. In voiced consonant the phoneme corresponding to /h/ does not exist and the fricative and the affricate are fused into one phoneme (e.g., /z/ and /dz/). Instead /r/ sound must be added. In the recognition of voiced consonant the difficulty arises in the presence of harmonic components during the noise part of fricative consonant and in the detection of consonant part of the stop consonant, because the noise energy is weak and the duration is short. The recognition of nasal consonant was not included in the system because the real time processing of the transient part needs complicated circuit.

The recognizer of connected speech of Fig. 3·1 processes some particular phoneme sequences, which are essential in connected speech. All the sequences must be processed considering the phonetic context, which, however, can not be realized in the system. The sequences processed in the block are long vowel, elision of vowel and double consonant. The long vowel sign is generated after the vowel phoneme when the duration of the vowel segment exceeds some threshold interval, though the context, the level of vowel and time pattern of pitch frequency are also the important parameters. The fricative or

affricate consonants such as /s/, /ʃ/, /ts/ and /tʃ/ are often pronounced without being followed by the vowel, when the vowel is to be followed by the stop consonant. In this case recognition between /s(u)/ and /ʃ(i)/, /ts(u)/ and /tʃ(i)/ must be carried from the properties of consonant itself without the knowledge about the following vowel, which was not attained in the system. As shown in Fig. 3·15 the distributions (and the spectra) of /s/ and /ʃ/, /ts/ and /tʃ/ resemble in Japanese, this problem may be solved in linguistic level utilizing the semantic informations. The double consonant is detected by the presence of fairly long silent interval between the preceding vowel or consonant and the following stop unvoiced consonant (refer to section 4·3 of PART I).

3·5 Combination of Recognition and Segmentation

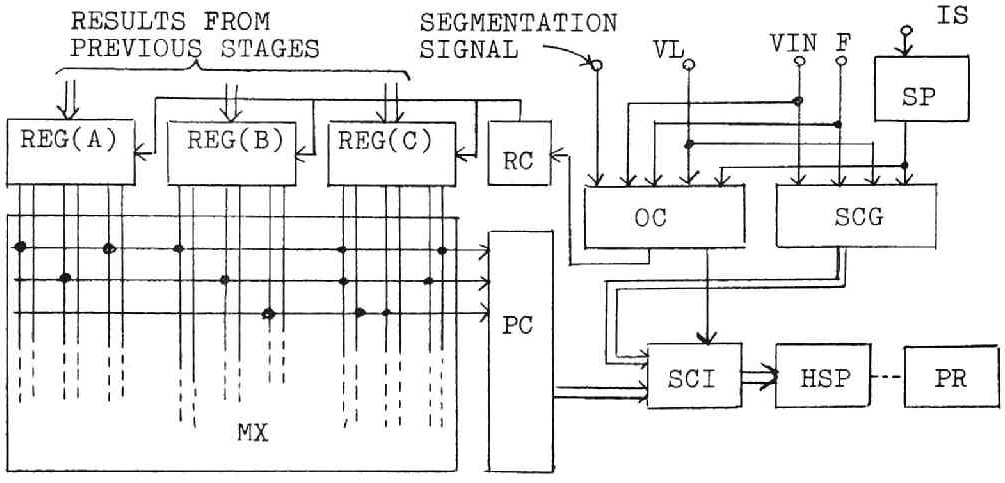
At the steps stated above, parameters and features of the input speech sound were obtained by the separate circuits, such as phoneme classifier, feature extactor, vowel recognizer, consonant analyzer, etc.. These results are obtained at various timings and manners. For instance the vowel circuit recognizes the each 20 ms segment during the vowel part, obtainging a series of vowel recognitions and on the other hand consonant analyzer, phoneme classifier, etc. operate in consonant segment, the results of which are obtained at various timings.

The function of phoneme recognition circuit (PHR) in Fig. 3·1 is to memorize these results and thereby to recognize the final output. Two types of output symbols are possible in Japanese; the Kana letter system and the Roman letter system. In the Roman letter system, a phoneme corresponds to one symbol and it seems that the recognition of consonant and vowel is accomplished separately. However, as stated before, it is necessary to take the mutual effect between consonant and vowel in consideration. This suggests that the Kana letter system, in which one symbol (letter) represent the phonemic constitution of consonant plus vowel (C+V), will be suitable as the output symbol.

The block diagram of phoneme recognition circuit and output circuit are shown in Fig. 3·17. The results obtained in previous stages are once stored in register REG(A), (B) and (C), each assigned to feature or phoneme group, vowel recognition series and binary signal of consonant analysis, respectively. The decision of the output, the Kana letter, is made in recognition matrix (MX), which is programmable diode matrix with 70 binary inputs (column) and 100 output symbols (row). The appropriate logics can be constructed by the plug in diodes. As stated in previous sections, the logics of consonant recognition are made in combination with the results of vowel recognition; that is, the logics for a particular consonant phoneme are different according to the following vowel. For instance, when it was known from phoneme classifier that the consonant belongs to unvoiced consonant without hiss, stop consonant and the duration was not long and from vowel recognizer that the following vowel was /a/, then the decision logics needed for consonant analysis (REG(C)) is the classification among /k(a)/, /t(a)/ and /p(a)/. (Refer to Fig. 3·13.) Puncher coder (PC) of Fig. 3·17 converts the symbol to a 8 bit code, which is led to high speed puncher (HSP) through special code inserter (SCI).

Fig. 3·18 is the time chart of phoneme recognition circuit for input sound with phoneme sequence $C+V_1+V_2$(C; consonant, V; vowel). The results of consonant analysis, phoneme classification, etc. obtained during the consonant segment are stored in registers as in (b) of the figure and the registers of vowel recognition are re-written successively each time the vowel sampling is made (e. g., every 20 ms). The segmentation signal which controls the timing of recognition is led from segmentation signal generator of Fig. 3·2. That is, the signal appears each time a new combination of the stabilities S_{ij} are detected in both F_1 and F_2 regions.

In Fig. 3·18 the time point (A) is the case when the stabilities are detected at the same time and (B) is the case when a stability has been already present in one of the F_1 and F_2 region and a new stability was found in the other region.



VL; long vowel, VIN; vowel insertion, F; double consonant, IS; speech sound period, REG(A)---(C); registers, MX; recognition matrix, PC; puncher coder, RC; reset circuit, OC; output control, SCG; special code generator, SCI; special code inserter, HSP; high speed puncher, PR; printer.

Fig. 3.17 Block diagram of phoneme recognition circuit and output circuit.

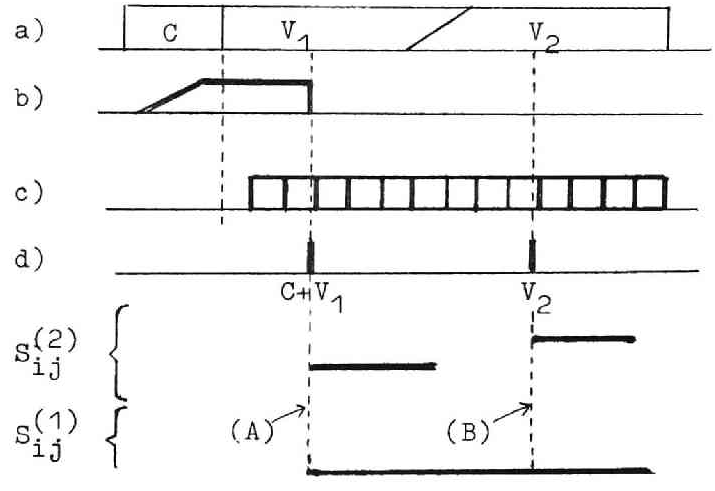


Fig. 3.18 Time chart of the phoneme recognition circuit. (a) Input speech(C; consonant, V; vowel). (b) Stored results of consonant analysis and phoneme classification. (c) Vowel recognition series. (d) Output control signal from segmentation part.

The phoneme recognition circuit combines, at the time when this segmentation signal is generated, the results of the consonant analysis, phoneme classifier already held in registers with those of the vowel recognizer received at that time. In the example of Fig. 3·18, one Kana letter for the combination of $C+V_1$ is at first printed out and next one Kana letter (vowel) for V_2. After one output has been sent out, all the registers are reset in preparation for the next operation. When the register (REG(A)) makes the indication of "vowel", a pure vowel symbol is printed.

The timing of recognition is given by output control circuit (OC). In normal phoneme sequences the signal is obtained from segmentation part. Some special symbols are prepared, such as long vowel (VL), vowel insertion (VIN) in case of elision of vowel and double consonant (F), which are processed in recognizer of connected speech, and space symbol (SP) which is put when silent interval is detected for more than 300 ms. The 8 bit codes of these symbols are combined in special code inserter with the codes from phoneme recognition circuit, the timings of which are designed so as not to over up. The code is punched on paper tape with the speed of 60 characters per second and afterward the paper tape is read and Kana letters are printed.

3·6 Conclusion

In this chapter a experimental model of the speech recognition system designed for Japanese speech sound was described. The system was at first designed for monosyllable recognition and afterward it was extended to accept some connected speech sounds by attaching the segmentation circuits. The recognition unit is the phoneme, so as not to limit the input category to a certain words. The segmentation of vowel sound to elementary vowel segments was tried by the detection of the "stability" from the zero-crossing pattern obtained by the circuits of chapter 2. The detection of the stability is effective for simple formant pattern but for complicated pattern spoken by fast articulation the detection was not successful,

which will be dealt in chapter 4, considering phonetic context. The principle of recognition procedure is the distinctive feature extraction, phoneme classification and the analysis of spectral features. From the mutual relationship and the time variations of envelopes of filters' outputs, feature detection and phoneme classification were performed. The operations are sensitive for level of the input speech sound, because of the logic based on absolute level. The number of filters must also be increased.

The analysis of vowel and consonant was made by zero-crossing analysis in separate circuit. The vowel was recognized in F_1 F_2 domain and the shifts of the formants by male and female voice were normalized by selecting the circuit constants by pitch frequency. Averaged score for male and female voices is about 94 %, though the circuits used are rather simple, for the better score some other principles such as filter analysis of chapter 3 of PART I, the zero-crossing analysis of chapter 2 of PART II, etc. must be utilized. The score of unvoiced consonants were about 70 % for clear articulation, which is largely affected by the method of articulation, though the score of each parameter used in recognition is better than this. For voiced consonants and nasals, the sufficient results were not obtained, because of the miss operation of sampling, the disturbance by lower components, etc..

The system was designed to process the speech sound in real time. Therfore simple methods and circuits were utilized. Though there have been left many problems unsolved, several circuits were developed for the parameter extractions. To solve the problems, much more complicated principles and circuits or computer simulation will be required.

Chapter 4

CONNECTED SPEECH RECOGNITION BY PHONETIC CONTEXTUAL APPROACH (9)(24)(33)

In the speech recognition system described above the speech sound was divided into several segments (that seem to correspond to the phoneme), from the time variation of the analyzed pattern and then discrimination was performed for each segment. The parameters or distributions of a segment are, however, affected by the phonetic contextual effect from the neighbouring phonemes. The speech sound, therefore, must be recognized as a whole pattern considering phonetic context. This principle, however, will be impossible to apply to the recognition system of the general conversational speech, because the number of words to be processed will become tremendous.

It will be valid simplification to consider that, though the reciprocal interactions will occur between the phonemes which are separately situated each other, the important influences to one phoneme will be the influences from the just preceding and the just following phonemes. As the general principle of phonetic contextual approach we selected, as the basic recognition unit (or segment), the segment of pattern corresponding to the three phoneme sequence.

As a preparatory step, trigram of the Japanese phoneme sequences was examined as described in chapter 5 of PART I, which gave us the data to design the recognition system. That is, though the possible number of combinations of the three phoneme sequences is too large, it was proved that the number of combinations actually used in daily conversation is reduced to the realizable scale. On this basis we adopted the conversational speech recognition system in which the phoneme was chosen as the recognition unit and the phonetic contextual effects from adjacent phonemes were considered. The principle was applied actually to the vowel recognition.

4·1 Principle of Recognition

The basic principle of recognition is the matching of the analyzed pattern of input speech sound with the stored standard patterns corresponding to the three phoneme sequences. The time variation of the analyzed pattern shows the different tendency according to the phoneme sequence that the speech sound contains. Further, it is affected largely by the articulating condition and individuality. Because of this variety the necessary number of the standard patterns prepared in the system will reach an unrealizable scale. As the effective method to solve this, we tried to express the pattern of speech sound by a set of the patterns; the sequence pattern and the weight pattern.

The sequence pattern is the pattern that shows the sequence of states, that is, how the parameters or the distributions of speech sound at a sampling point change with the time. The weight pattern is a series of numbers that show the number of the successive occurrence of each state which compose the sequence pattern. This expression of pattern may be useful to any pattern other than speech pattern, but it works more effectively to the pattern of speech sound where the parameters' changes are negligible or gradual in the most portion of sound except for some time points. The effect of the speed of articulation on the pattern is in this method expressed mainly in the size of the weight pattern. The sequence pattern and the shape of the weight pattern are not influenced largely by the articulation speed.

In the system the matching operation is first executed in each of the sequence pattern and the weight pattern in parallel and then both results are combined to produce final matching of the whole input pattern. This two-stage logic of matching reduces the complexity of logic and circuit. By this expression the required memory capacity of the equipment is also greatly saved as well as the simplification of operation.

4.2 System of Recognition Circuit

Fig. 4·1 is the block diagram of the system. The timings of the whole system and that of the input pattern are synchronized by the shift-pulse timing. A new input coming to the input terminal is compiled either to the weight

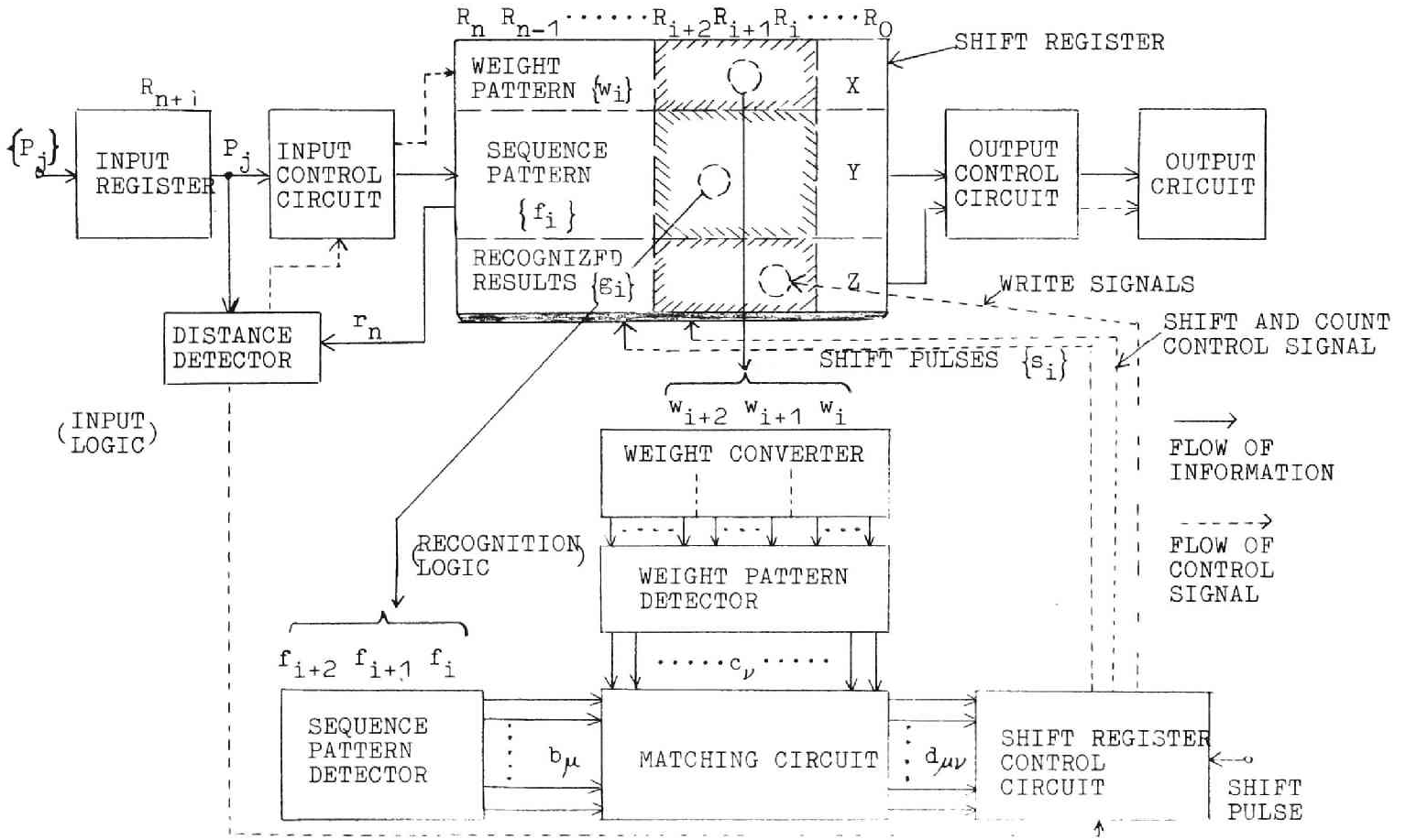


Fig. 4.1 Block diagram of speech recognition system, considering the phonetic contextual effect.

pattern or to the sequence pattern in the shift register. During the information flows from the input stage R_n to the output stage R_0 of the shift register memory, logical operations are successively carried on its contents. Thus, the memory capacity of the shift register needs not be so large and in principle there is no limitation in the length of the input speech sound.

In the processing of speech pattern expressed as the sequence of states, sequential logic circuit is necessary. However, the speech pattern is not perfectly sequential from the top of one connected speech sound to the last part, but a certain state is related only with its neighboring states. The initial state of the sound pattern does not effect to the later part apart from it. In the shift register the pattern is expressed as space pattern, on which the combinational logic circuit performs the "running" sequential logic. (It may be called "running" in association with the running spectrum obtained by band pass filter, because any state of pattern can be an initial state of sub-sequence.)

Another feature in the processing of speech recognition is that the present state is influenced by the following state to come as well as the preceding state, being different from the physical phenomenon. By processing the pattern on shift register, these features are well treated in real time.

Any sequential circuit has inputs, memory and the combinational logical circuits connected to them. The next state of memory is decided by the present state, the next input and the logics of circuits. In this pattern recognizer memory consists of shift register and the input is fed from its input stage. As a principle the state of shift register is shifted by one, if no logical modification is exerted. This constraint is very strong and particular. Two logical circuits are prepared: The input logic circuit compiles the input pattern to sequence pattern and weight pattern and recognition logic circuit processes the pattern stored in shift register. The modifications are to re-write the states of X, Y and Z part and to eliminate the redundunt stages.

The logical circuits, which get the input from the shift register memory,

are designed to work in parallel (static logic). After the shift register changed the state by the shift pulse, during the time interval till the next shift pulse is applied the shift register, the logical circuits and the shift register control circuits are kept in the constant state. By the next shift pulse the new input is fed to the shift register and at the same time the shift register control circuit puts its output logic into practice to control the shift register and the output circuit. The processing of the input pattern is, therefore, continued in real time as pattern flows from the input to the output.

Let us consider the speech pattern P as the time series of the parameters or distributions P_j obtained in the time interval j. Then P is written as

$$P = \{ P_j \} = P_1, P_2, \ldots\ldots\ldots, P_j, \ldots\ldots$$

We define the word "run" as follows; a longest series of the parameters taken from the pattern $\{P_j\}$, where all the adjacent paramaters are recognized as the same based on some criteria. The sequence of runs obtained from the parameters $\{P_j\}$is denoted by $f_1, f_2, \cdots\cdots, f_k, \cdots\cdots$, where it is supposed that $P_{j-1} \quad f_k$, and the length of each run denoted by w_k. Then a set of $\{f_k\}$ and $\{w_k\}$ have the one to one correspondence with the original pattern $\{P_j\}$ and they also correspond to the sequence pattern and weight pattern, respectively, just we have defined them above.

In real system, P_j is given as a time series. In the j-th time interval t_j between the time points j and j 1, the latest part of the input pattern is stored as $\{f_k\}$, $\{w_k\}$in the shift register and the next input P_j exists in the input register. Distance detector of Fig. 4·1 computes the distance between the contents of the input register R_{n+1} and that of the R_n stage of the shift register. Input control circuit compiles, under the control of the distance detector, a new input in R_{n+1} to the pattern which is expressed as $\{f_k\}$, $\{w_k\}$ in the shift register.

Now let r_{nj} be the contents of the register R_n in the time interval t_j, then $r_{n+1j} = P_j$. On the other hand $P_{j-1} \quad f_k$, then $r_{nj} = (w_k, f_k)$ and $r_{n-1\ j} \quad (w_{k-1}, f_{k-1})$.

The logics in distance detector are as follows.

(a) For a given threshold ε, if the distance q_j /P_j $f_k/\leq\varepsilon$, then it is recognized as $P_j = f_k$, and at the next time interval t_{j+1}, R_n is changed to $r_{nj+1} = (w_k + 1, f_k)$ without any change in the other part of the shift register.

(b) If $q_j > \varepsilon$, then it is recognized as $P_j \neq f_k$, and at the next time interval t_{j+1} all the contents of shift register are shifted by one to the output side and the R_n is set as $r_{nj+1} = (1, f_{k+1})$, where P_j f_{k+1}. (there are some subsidiary functions omitted here.)

The pattern stored in the shift register is detected by the diode logic of the detection circuits which are connected to stages R_i, R_{i+1} and R_{i+2}. That is, the length of pattern to be looked up at one time is the length of three runs. The sequence patterns (f_i, f_{i+1}, f_{i+2}) are led to the sequence pattern detector, where one of the standard patterns $\{b_\mu\}$ which are set as diode logic in the circuit is selected. A sequence pattern is considered to correspond to more than one phoneme sequence or it may correspond to a transition part. Which phoneme sequence is the desired output is decided by the detection of the weight pattern.

The weight w_i which is stored in shift register as a count number is converted to the relation of inequality by weight converter. This circuit may be considered as a decoder that generates, for each of the weights w_i, w_{i+1} and w_{i+2}, a number of outputs $\{a_{i\,\ell_1}\}$, $\{a_{i+1\,\ell_2}\}$ and $\{a_{i+2\,\ell_3}\}$ by the following logics.

When some value ℓ is given for w_i, the logic of $a_{i\ell}$ is

$$\begin{aligned} a_{i\ell} &= 1 \qquad \text{when } w_i < \ell \\ a_{i\ell} &= 0 \qquad \text{when } w_i \geq \ell \end{aligned} \tag{4.1}$$

and there are also the outputs of its negation $\{\overline{a_{i\ell}}\}$

Weight pattern detector decodes the weight pattern as a combination of the inequality relations, and gives signals to some of the prepared channels $\{c_\nu\}$. For example if it is required to generate the signal for the weight

pattern

$$\ell'_1 \leq w_i < \ell_1, \quad \ell'_2 \leq w_{i+1} < \ell_2, \quad \ell'_3 \leq w_{i+2} < \ell_3 \qquad (4\cdot2)$$

then the decoding logic is

$$c_\nu \quad a_i \ell_1 \cdot \overline{a_i \ell'_1} \cdot a_{i+1} \ell_2 \cdot \overline{a_{i+2} \ell'_2} \cdot a_{i+2} \ell_3 \cdot \overline{a_{i+2} \ell'_3} \qquad (4\cdot3)$$

The matching circuit combines the detected sequence pattern b_μ with some weight pattern c_ν's that are necessary to distinguish the possible phoneme sequences from the detected pattern b_μ. Generally there are several weight patterns to be combined with a certain sequence pattern. The logic among such c_ν's are exclusive, but not always exclusive between the c_ν's which are not combined to the same sequence pattern. The results of the matching are the decision for the run of length three, that is, a part of the whole input pattern $\{P_j\}$. By repeating this procedure successively in each time interval, the final recognition result is obtained.

After a shift pulse was applied, one of the output lines $d_{\mu\nu}$ is being selected. The state of the shift register at the next time interval is decided by the logics of the shift register control circuit selected by $d_{\mu\nu}$. The function of shift register control circuit is to control the shift pulse of each stage (a shift pulse from R_n to R_{n-1} is written as s_{n-1}) and to write the recognized results. The shift pulse at each stage is independently controlled by the logic of this circuit which is determined by the selected input line $d_{\mu\nu}$. That is, if it is required to eliminate the contents of the R_{i+1}, then the necessary logics are

$$s_0 = s_1 \quad \cdots\cdots = s_i = 0 \text{ and } s_{i+1} \quad s_{i+2} = \cdots\cdots \quad s_n = 1 \qquad (4\cdot4)$$

($s_i = 0$ means that no shift pulse is added to R_i.)

The elimination procedure occurs when the matched part of the pattern is the transient segment from one phoneme to the next. On the contrary, when the part was recognized as a recognition unit(phoneme), write signal puts the result into the shift register stages corresponding to the pattern under processing.

After the processing of the pattern is finished, it flows out of the output side R_0 of the register. The parts of the pattern recognized as the

recognition unit are picked up by the output control circuit to be sent to the output circuit.

4.3 Experiment of Vowel Pattern Recognition

The application of the above principle to the conversational speech recognition system will need considerably large scale devices. The system was applied to vowel pattern recognizer, because the structure of vowel is rather simple than consonant and the phonetic context in vowel segment is the problem that must be first solved in conversational speech recognition system. The vowel pattern recognizer described in this chapter was combined with the speech recognition system stated in the previous chapter by replacing the function of vowel segmentation circuit. The block diagram of the new system is shown in Fig. 4·2. The vowel pattern recognizer is inserted between vowel series recognizer, which was called as vowel recognizer (VR) in chapter 3, and phoneme recognition circuit and recognizes the vowel phonemes, based on the vowel sequence sampled and recognized each 20 ms as shown in Fig. 3·3. When the presence of a phoneme was detected the signal controls the operation of output through output control circuit (OC), on behalf of the segmentation signal. At the same time, the phonemic information enters the vowel register (REG B) of phoneme recognition circuit, in which final decision of letters is carried out. The operations of the vowel pattern recognizer, the vowel analyzer and the vowel series recognizer are synchronized by the clock pulse.

As in vowel segment the movements of formants are gradual, the pattern representation by sequence pattern and weight pattern is effective. The pattern of formants will be desirable as the input, but the number of patterns to be processed will become too large to construct the special purpose machine. The pattern used in the experiment is the recognition sereis of five vowels obtained in vowel series recognizer for every 20 ms.

The shift register is composed of 6 stages. The weight part (X) composed of 4 bit counter memorizes weight pattern $\{w_i\}$. The sequence part (Y) which

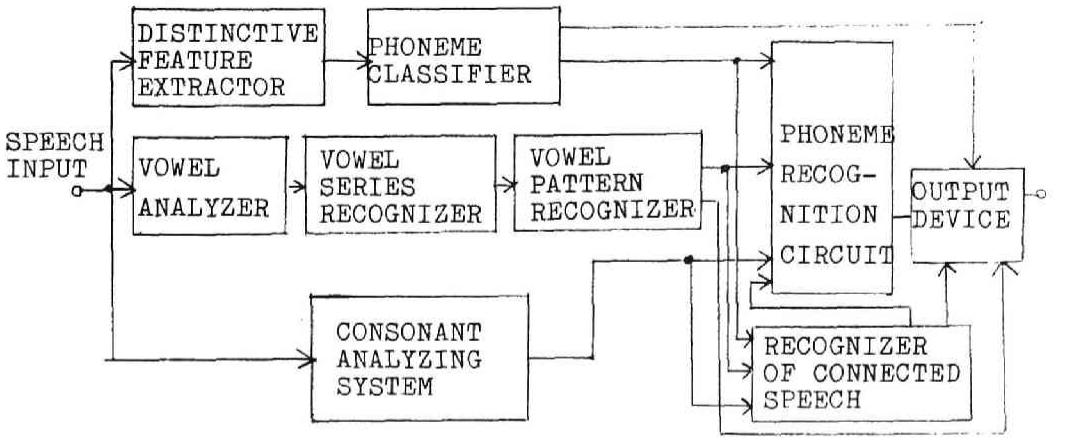
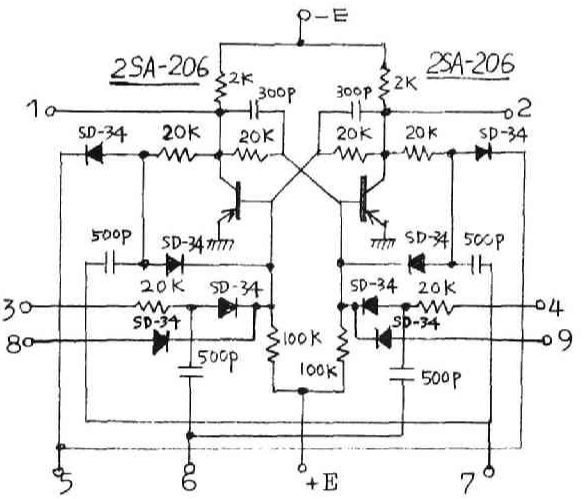
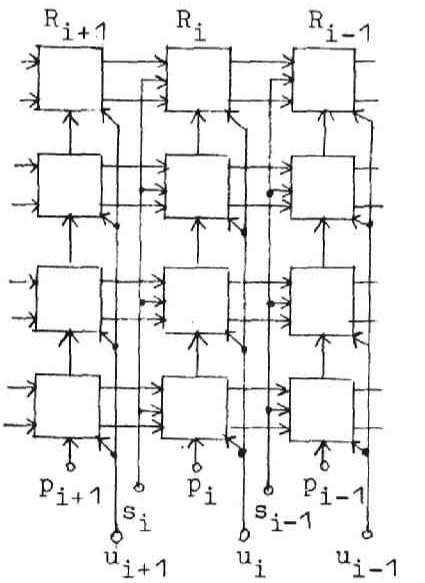


Fig. 4.2 Block diagram of speech recognition system combined with vowel pattern recognizer.



1,2; Outputs of circuit
3,4; Inputs for shift register (from previous stage)
5; Shift—count control signal
6; Shift pulse input
7; Count pulse input
8, 9; Set and reset inputs

(a) Circuit diagram of shift register unit.



p_i; Count pulse
s_i; Shift pulse
u_i; Shift-count control

(b) Connection of unit circuits of weight part(X) of the shift register.

Fig. 4.3 Circuit and the connection of shift register unit circuit.

holds sequence pattern $\{f_i\}$ is composed of 5 bits, each assigned for vowel "a", "i", "u", "e" and "o". (Coding was not tried.) The recognized results part(Z) of 4 bits is used to register the recongized results from recognition logic circuit.

Fig. 4·3 shows the unit circuit of shift register and its connection in weight part. In weight part the units must work as counter for the counting up of the length of each run. On the other hand when a new run is detected, the contents must be shifted to the next stage. Therefore, the vertical connection of Fig. 4.3 forms the counter and the horizontal connection forms the shift register, the selection of which is performed by shift-count control signal applied to each unit from distance detector and shift register control circuit. On operating as shift register, the shift-count control signal is pulled down to negative voltage to inhibit the counting up. The maximum weight is 16, i. e., 320 ms for 20 ms sampling. When the length of input run exceeds this value, the counter overflows and keeps the maximum value until it is shifted to the next stage. The contents of each stage of the sequence part (X) and the recognized part (Z) can be re-written by the write signal obtained from shift register control circuit, according to the recognition logics.

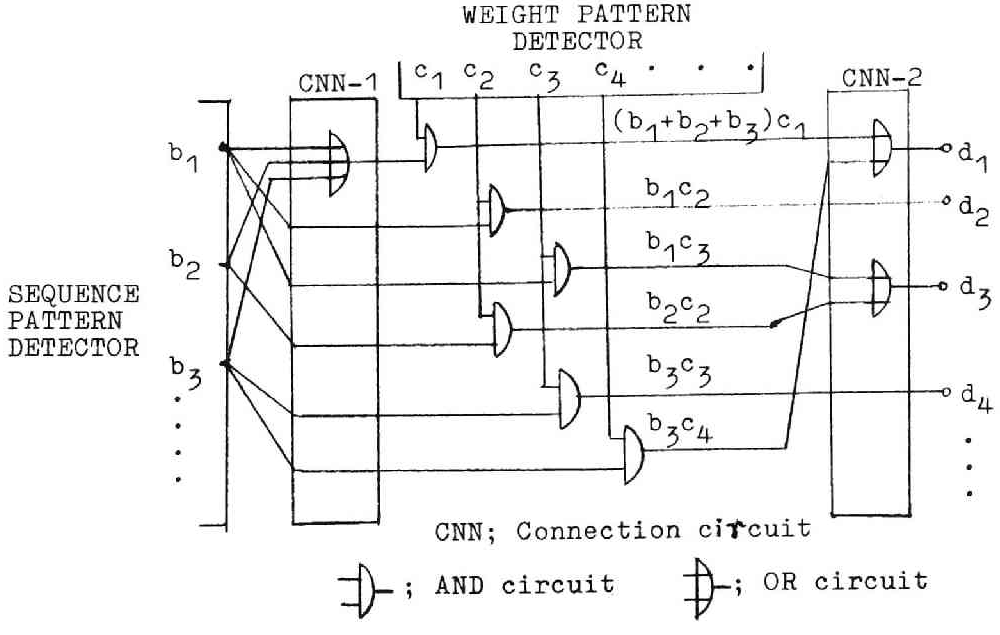
An example of operation of the distance detector is shown in Fig. 4·4 (a). When y_5 and y_4 (y_5 and y_4 mean the contents of sequence part of R_5 and R_4, respectively) are equal, y_5 is eliminated and compiled into R_4 at the next time interval, in such way as the weight of R_4 is increased by one and the other stages remain unchanged. When y_5 and y_4 are not equal, r_5 is shifted to R_4, r_4 to R_3, etc.. As seen in the state at timing t_4 of Fig. 4·4(a) in the case that y_4 and r_6 are the same but differ from y_5, y_5 is regarded as a noise component and at the next step it is eliminated and the weight of R_4 is increased by one. In summarizing, the logics are:

	R_6	R_5	R_4	R_3
t_1	a_1	·	·	·
t_2	a_2	a_1	·	·
t_3	e	a_2	a_1	·
t_4	a_3	e	$2a_1$	·
t_5	i_1	a_3	$3a_1$	·
t_6	i_2	i_1	$4a_1$	·
t_7	·	i_2	i_1	$4a_1$
t_8	·	·	$2i_1$	$4a_1$

(a) An example of input logic for input series "a_1 a_2 e a_3 i_1 i_2 ····". $2a_1$ means that the weight of a_1 is 2.

	R_5	R_4	R_3	R_2	R_1	R_0
(a)		V_4	V_3	V_2		
(b)			V_3	V_2	V_1	
(c)				V_2	V_1	*
(d)					V_1	*

(b) Connection of recognition logic circuits to shift register.
(a), (b), (c) and (d) are the pattern class explained in the text. The framed parts are connected to the circuits. V is the vowel state. * is the state in which no vowel state is detected.



(c) An example of connection of matching circuit.

Fig. 4.4 Operation and constitution of vowel pattern recognizer.

$s_5 = 1$, $s_4 = \cdots\cdots\cdots = s_0 \quad 0$, $u_4 \quad 1$

for $y_5 = y_4$

and for $y_5 \neq y_4 = r_6$, and

$s_5 = \cdots\cdots\cdots\cdots = s_0 = 1$, $u_4 = 0$

for $y_5 \neq y_4 \neq r_6$,

in which $u_4 \quad 1$ means to count up by one at the stage 4.

The state of sequence pattern stored in Y part is detected by sequence pattern detector. The sequence pattern of vowel part can be classified into several groups;(a) transient pattern, (b) phonemic pattern, (c) and (d) pattern with length 2 or 1. (a) is the pattern that may possibly represent transient part from one phoneme to the next. In this case the decision whether the detected pattern corresponds to transient part or to be recognized as one phoneme, must be executed, according to the corresponding weight pattern. For instance, for phoneme sequence /e-u/, the sequence pattern obtained by analysis may be "e o u ", in which "o" is the transient state from /e/ to /u/ and it must be eliminated. However, the articulation of /e-o-u/ is also possible, in which case state "o" must be recognized as one phoneme /o/. (b) is the pattern that is recognized as one vowel phoneme, in which the presence of vowel phoneme is expected by the movements of formants. For example, in the sequence "e a o", the first formant shows convex curve which is by no means a transient part, but shows the presence of /a/. In some cases the distinction of semi-vowels (/w/ and /j/) is needed; for example, "a e a" may be /a-e-a/ or /a-j-a/ and "i e a" may be /i j a/ or /j a /. These patterns of (a) and (b) are composed of three states, the shorter patterns are processed in (c) and (d) groups.

The connection of these patterns to the shift register is shown in Fig. 4.4 (b). The transient pattern (a) is processed in earlier step and the redundunt state is eliminated. After such contraction, the detection

of phonemic pattern (b) is carried out. The patterns set in sequence pattern detector $\{b_\mu\}$ are about 80, each of which is combined with the respective weight patterns in matching circuit.

The weight patterns to be matched to the sequence pattern of Fig. 4·4 (b) are obtained from weight pattern w_i , X part of the shift register. Weight converter, prepared for each stage of R_1, R_2, R_3 and R_4, converts each weight to a set of logical variables $\{a_{i\ell}\}$ in such way:

$$a_{i\ell} = 1 \quad \text{for} \quad w_i \geqslant 1 \quad \text{and} \quad a_{i\ell'} \quad 1 \quad \text{for} \quad w_i < \ell'$$

in which: $\ell = 2,3,4,5,8,10,13.$

ℓ' $2,3,4,5,8,10.$

"1" means one sampling interval of 20 ms.

The circuit is composed of diode logical circuit and generates 50 variables.

In the weight pattern detector logics of (4·3) are set by diode logics. The number of outputs $\{c_\mu\}$ is 45, which are combined with the sequence patterns. Matching circuit is composed of 55 two input diode gates, the inputs of which are connected to sequence pattern detector and weight pattern detector. A schematic example of maching circuit is shown in Fig. 4·4(c). To merge the logics, connection circuit(CNN), OR gates, are used. The outputs of matching circuit, whose logics in shift register control circuit are the same, are merged to d_1, d_2, $\cdots\cdots d_{25}$, by CNN-2.

The shift register control circuit is composed of diode logical circuit and performs the shift pulse control, shift count select and writing control of each stage. As stated above, when contents of i-th stage are to be eliminated, the shift pulse s_{i-1} is not generated and interlockingly X_{i-1} of the shift register is switched so as to constitute the counter connection, by which the counter is increased by one. As the writing of the recognition results to Z part of the shift register is drived by shift pulse, no writing is performed at that time interval, if the shift pulse to that stage is stopped. Therefore, if shift of pattern is stopped by the input logic circuit, all the operations detected in shift register control circuit are not executed.

Marks are put on Z part of shift register, when vowels, long vowel and semi-vowels (/i/ and /w/) were recognized. For instance, if the state in R_3 was recognized as vowel phoneme, vowel mark is put on the state. If a sequence of states in R_3 and R_4 were recognized as /ja/ vowel mark and semi-vowel mark are put on R_4(i. e., on state "a").

The normal period of shift pulse is 20 ms, by which all the circuits of vowel recognition system are synchronized. The minimum time delay till the recognized pattern appears to the output side of the shift register is 120 ms. Above that, while the states of input sequence remains the same as in long vowel, the pattern may not appear to output side and 120 ms are needed to reach to it after the end of vowel segment was detected. It may possibly occur that the next consonant or vowel comes, before the preceding vowel pattern remains in shift register. This causes some confusions, because the informations of consonant and phoneme classification, stored in phoneme recognition circuit of Fig. 4·2 and to be combined with the results of vowel recognition, must be reset before the next consonant or syllable comes. To avoid this difficulty, when the space —— no existence of vowel state —— was detected at R_6 or at any stage of shift register, the output side of the shift register is shifted by faster timing of 1.25 ms, being separated from the input side. When a letter to be punched comes to output side the register is again swithed to the normal mode.

The output of Z part is connected to the output device of Fig. 4·2. The mark of vowel and long vowel controls the output operation as explained in the previous chapter.

The operations of the circuit are explained in Table 4·1(a) and (b) for input word /k a u / and /s a j o:/, respectively. Of course the consonants /k/ and /s/ are omitted in the examples. The state of each step of register is expressed as 7u(V), etc., which means the weight of state "u" is 7 and it was recognized as vowel (V). The operations of shift register are controlled by the input logic and the recognition logic. In normal

Table 4.1 Examples of operation of vowel pattern recognizer.

(a) For word /ka-u/ (The input series is "6a,3o,7u")

TIMING NO.	CLOCK MODE	LOGIC NO. DETECTED			R_6	R_5	R_4	R_3	R_2	R_1	R_0	PH
		B	C	D								
1	N		IN		o	a	5a		-	-	-	
2	N		SF		o	o	6a	-	-	-		
⋮												
5	N		IN		u	u	3o	6a	-	-		
6	N		SF		u	u	u	3o	6a	-	-	
⋮												
10	N		IN		u	u	5u	3o	6a	-	-	
11	F		IN		-	u	6u	3o	6a	-		
12	F	1	16	3	-	-	7u	3o	6a	-	-	
13	F		SF		-	-	-	7u	6a	-	-	
14	F	2	9,10	4,5	-	-	-	-	7u	6a	-	
15	F		SF		-	-	-	-	-	7u(V)	6a(V)	
16	N		SF		-	-	-	-	-	-	7u(V)	/a/
17	N		SF		-	-	-	-	-	-	-	/u/
18	F	(END)										

Notes; (a) "6u" shows the state "u" has the weight 6.
(b) "-" means that no vowel is detected.
(c) CLOCK MODE; "N" is normal mode of 20 ms. "F" is fast mode of 1.25 ms.
(d) LOGIC NO. DETECTED; "SF" means that all the contents of shift register are shifted by one. "IN" means that the shift register is controlled by input logic. "B","C" and "D" are the identifier of the logics, representing sequence pattern, weight pattern and shift register control circuit, respectively. The numerals are the numbers of detected logics as shown in logic table. (Table 4.1(c))
(e) The mark set in Z-part is shown by ( ), in such way as 7u(V), 6a(V), etc., in which V, L, Y, and W represent vowel, long vowel, /j+vowel/ syllable and /w+vowel/ syllable, respectively.
(f) PH means the output phoneme to be printed .

Table 4.1(b) For word /sa-jo:/
The input series is "4a,o,2e,3i,e,8o".

STEP NO.	CLOCK MODE	LOGIC NO. DETECTED			R_6	R_5	R_4	R_3	R_2	R_1	R_0	PH
		B	C	D								
1	N		SF		e	o	4a		–	–	–	
2	N		SF		e	e	o	4a	–	–	–	
3	N		IN		i	e	e	o	4a	–	–	
4	N	3	5	3	i	i	2e	o	4a	–	–	
5	N		IN		i	i	i	2e	4a	–	–	
6	N		IN		e	i	2i	2e	4a	–	–	
7	N	4	7	3	o	e	3i	2e	4a	–	–	
8	N		SF		o	o	e	3i	4a	–	–	
9	N		IN		o	o	o	e	3i	4a	–	
⋮											–	
15	F		IN		–	o	7o	e	3i	4a		
16	F	5	7	3	–	–	8o	e	3i	4a	–	
17	F	6	19	12	–	–	–	8o	3i	4a	–	
18	F	2	9,10	4,5	–	–	–	–	8o (Y,L)	4a	–	
19	F		SF		–	–	–	–	–	8o (Y,L)	4a (V)	
20	N		SF		–	–	–	–	–	–	8o (Y,L)	/a/
21	N		SF		–	–	–	–	–	–	–	/jo:/
22	F		(END)									

Table 4.1(c) List of logics used in the examples. Only the logics necessary for this illustration are picked up in this table from the logics set in the device.

SEQUENCE PATTERN (B)						WEIGHT PATTERN (C)					LOGIC OF SHIFT REGISTER CONTROL(D)		
NO.	f_4	f_3	f_2	f_1	f_0	NO.	w_4	w_3	w_2	w_1	NO.		
1	a	o	u	-	-	8	≥ 3	≥ 4	≥ 2	-	1	Y_4, V_3, V_2	
						15	≥ 3	<4	<2	-	2	Y_4	
						16	≥ 3	<4	≥ 2	-	3		3/
2	-	-	Ⓥ	a	*	9	-	-	-	≥ 3	4	V_1	
						10	-	-	≥ 3	-	5	V_2	
						11	-	-	-	≥ 13	6	L_1	
						12	-	-	≥ 13	-	7	L_2	
3	e	o	a	-	-	1	<3	<2	<2	-	3		
						2	<3	2	<2	-	8		4/
						3	≥ 3	<3	<2	-	3		
						4	<3	≥ 3	<2	-	9	V_3	
						5	<3	<3	≥ 2	-	3		
						6	<3	≥ 3	≥ 2	-	8		
						7	≥ 3	<4	≥ 2	-	3		
						8	≥ 3	≥ 4	≥ 2	-	9		
4	i	e	a	-	-	THE SAME AS B-3							
5	o	e	i	-	-	1,3,7					3		
						2					8		
						4, 8					9		
						13	<3	$2\langle,4$	≥ 2	-	8		
						14	<3	≥ 4	≥ 2	-	10	V_3	4/
6	-	o	i	Ⓥ	-	17	-	≥ 3	≥ 2	≥ 2	11	Y_3	
						18	-	≥ 3	≥ 10	≥ 2	5		
						19	-	≥ 8	≥ 2	≥ 2	12	Y_3, L_3	

Notes; -; No connection with shift register
*; State other than vowel.
Ⓥ; Any vowel state.
>3; means that the weight is greater than 3
Y_3 etc.; mean to write the recognized results to the stage indicated by the suffix.
3/ etc.; mean to eliminate the stage designated.

condition shift register is shifted by one step (denoted by SHIFT in the Table) and when input logic works (denoted by INPUT), logics detected in recognition logic circuit are neglected. The detected logics of recognition logic circuit are presented by identifying number such as B-1, C-1, D-1, etc.. The logics used in the examples are shown in Table 4·1 (c), which were extracted from the logics set in device. In some cases the detected logic of C is not unique, for instance in step No. 14 of Table 4·1 (a), and accordingly two logical circuits are excited at shift register control circuit by D-4 and D-5. However, the resultant operation is not contradictory, that is, the simultaneous set of two vowel marks at stage 2 and 1. In the same way, in step No. 18 of Table 4·1 (b) recognitions of /j+vowel/ and long vowel are performed.

The speech recognition system of Fig. 4·2 using this vowel pattern recognizer can process the spoken word, though there are a few limitation in input category and in the speed of articulation at present stage. The operation of the vowel recognizer is better than the method using the stability. The faults in judgment are mainly caused by the error in the input vowel series, because the input to vowel pattern recognizer is simplified to the five vowels.

4·4 Conclusion

In this chapter the principle of speech pattern recognition and its application to vowel pattern recognition system were described. The three phoneme sequence was selected as a recognition unit to do pattern matching considering the phonetic context. (The possibility of the realization of this method was discussed in chapter 5 of PART I.) The speech pattern was presented by a combination of the sequence pattern and the weight pattern in shift register memory, which served for the effective processing of pattern matching as well as for the reduction of the memory capacity needed for the processing. The sequence pattern and the weight pattern are

processed in real time during the signal flows on the shift register memory, therefore there is no limitation in the length of the input sound. According to this principle vowel pattern recognizer was devised which, accepting the vowel recognition series, performs the segmentation to the phoneme segments and the recognition of the segments at the same time. By the expression of pattern as a combination of sequence pattern and weight pattern, the time scale of the matching logics were presented in relative relation and, therefore, the logics are independent of the speed of articulation, unless it is not so fast that the pattern is deformed appreciably by co-articulation effect. The vowel pattern recognizer could operate for the connected vowel including two or three vowel phonemes.

The vowel recognizer was combined to the speech recognition system stated in chapter 3, which could work for short words containing the connection of two- or three-vowel phoneme sequence and a semi-vowel, such as /kau/, /kai/, /teoke/, /aoi-ie/, /tsuja/, /sajo:/, etc..

Chapter 5

CONCLUSION

In PART II an automatic recognition system of the Japanese speech sounds and the zero-crossing analysis method were described. An automatic recognition system that operates in real time for the Japanese connected speech sounds was described in chapter 3, using the zero-crossing analysis method stated in chapter 2. The pattern recognition system by means of the phonetic context was explained in chapter 4, which was combined with the recognition system of chapter 3.

In chapter 2 the descriptions were made on the representation of the zero-crossing wave, the analysis circuit of zero-crossing intervals and zero-crossing analysis of formant signals and speech sounds. A notion of phase sampling was proposed, by which the zero-crossing wave is considered as the sampling of the signal at every π rad. It was shown that the improvement of the articulation score of zero-crossing wave by differentiation has relation with the spectrum of the zero-crossing wave. Zero-crossing analysis circuit was devised, which measures the zero-crossing intervals by digital method and obtains the zero-crossing distribution expressed by pulse numbers and then expressed by a set of voltages, operating successively under the control of periodic sampling signal. This circuit was used as the analysis circuit in the speech recognition system of chapter 3 and chapter 4.

Results of zero-crossing analysis suffer much influence from the characteristics of filter used before the conversion to zero-crossing wave. In this chapter the analysis of formant structure was performed by passing the signal through single tuned filter. The sequence of zero-crossing intervals of the signal with one formant, passed through single tuned filter, was found to vary largely according to the relation of band widths of both reasonant circuits. Some methods were proposed that extract the formant structure of the speech sound by the analysis of zero-crossing waves passed

through single tuned filters. The method was adopted that obtains so called "summed zero-crossing distribution" from a set of the zero-crossing distributions stated above. The formant structure at the onset of vowel was compared with that of the burst of unvoiced consonant, by which the differences of those signals were presented. The method was also applied to the analysis of noise sounds having stationary noise at high frequency region, such as /s/, /ʃ/, /h(i)/, /ts/, /tʃ/ and /k(i)/, and could detect the lower formants (the second or the third) which are often masked by higher frequency components. It was deduced from the results that the behaviors of such formants are important in the distinction of the noise consonants.

In chapter 3 an experimental model of the speech recognition system designed for the Japanese speech sound was described. The system was at first designed for monosyllable recognition and afterward it was extended to accept some connected speech sounds by attaching the segmentation circuits. The recognition unit is the phoneme, so as not to limit the input category to a certain words. The segmentation of vowel sound to elementary vowel segments was tried by the detection of the "stability" from the zero-crossing pattern obtained by the circuits of chapter 2. The detection of the stability is effective for simple formant pattern but for complicated pattern spoken by fast articulation the detection was not successful, which will be dealt in chapter 4, considering the phonetic context. The principle of recognition procedure is the distinctive feature extraction, phoneme classification and the analysis of spectral features. From the mutual relationship and the time variations of envelopes of filters' outputs, feature detection and phoneme classification were performed. The operations are sensitive for level of the input speech sound, because of the logic based on absolute level. The number of filters must also be increased.

The analysis of vowel and consonant was made by zero-crossing analysis in separate circuit. The vowel was recognized in F_1 F_2 domain and the

shifts of the formants by male and female voice were normalized by selecting the circuit constants by pitch frequency. Averaged score for male and female voices is about 94 %, though the circuits used are rather simple. For the better score some other principles such as filter analysis of chapter 3 of PART I, the zero-crossing analysis of chapter 2 of PART II, etc. must be utilized. The score of unvoiced consonants were about 70 % for clear articulation, which is largely affected by the method of articulation, though the score of each parameter used in recognition is better than this. For voiced consonants and nasals, the sufficient results were not obtained, because of the miss operation of sampling, the disturbance by lower components, etc..

In chapter 4 the principle of speech pattern recognition and its application to vowel pattern recognition system were described. The three phoneme sequence was selected as a recognition unit to do pattern matching considering the phonetic context. The speech pattern was presented by a combination of the sequence pattern and the weight pattern in shift register memory, which served for the effective processing of pattern matching as well as for the reduction of the memory capacity needed for the processing. The sequence pattern and the weight pattern are processed in real time during the signal flows on the shift register memory, therefore there is no limitation in the length of the input sound. According to this principle vowel pattern recognizer was devised which, accepting the vowel recognition series, performs the segmentation to the phoneme segments and the recognition of the segments at the same time. By the expression of pattern as a combination of sequence pattern and weight pattern, the time scale of the matching logics were presented in relative relation and, therefore, the logics are independent of the speed of articulation, unless it is not so fast that the pattern is deformed appreciably by co-articulation effect. The vowel pattern recognizer could operate for the connected vowel

including two or three vowel phonemes.

The vowel recognizer was combined to the speech recognition system stated in chapter 3, which could work for short words containing the connection of two- or three-vowel phoneme sequence and a semi-vowel.

The system was designed to process the speech sound in real time. Therefore simple methods and circuits were utilized. One of the purposes of this recognition system is to find the processing method for the extraction of speech parameters. It is also necessary to find a general principle to process the speech from the phonetic contextual point of view. The combined expression of the pattern by the sequence pattern and the weight pattern may be one of the effective ways. Though there have been left many problems unsolved in this system, several circuits were developed for the parameter extractions. To solve the problems, much more complicated principles and circuits or computer simulation will be required.

ACKNOWLEDGEMENTS

I wish to express sincere thanks to Professor Toshiyuki Sakai for his guidance and support during my course of study.

I also wish to express sincere thanks to Professor Ken-ichi Maeda for his encouragement and support to this work.

I am indebted to Dr. S. Inoue and Mr. K. Kagiyama for their helpful suggestions given on starting my work, to Mr. T. Kurashita, Mr. K. Shirai, Mr. K. Hashimoto, Mr. M. Nagao, Mr. Y. Ni-imi and Mr. K. Tabata for their co-operations and helps in experiments.

I am indebted to Mr. H. Nishio and Mr. M. Nagao for their suggestions given in daily discussions on automatic recognition.

I am also indebted to Dr. T. Kurokawa, Dr. H. Tomonari, Mr. T. Tsuzaki, Dr. T. Sekimoto, Dr. K. Nagata, Mr. Y. Kato, Mr. H. Kaneko and Mr. H. Kondo (the late) of Nippon Electric Company Ltd. (NEC), Tokyo, Japan, for the realization of the devices of the recognition system.

REFERENCES OF PART I

(1) Gunnar Fant,"Acoustic Theory of Speech Production," Mouton & Co. 1960 's-Gravenhage.

(2) R. R. Rietz and L. O. Schott, "The Visible Speech Cathode-Ray Translator," JASA 18, 50, 1946.

(3) F. Vilbig and K. H. Hause, "Some Systems for Speech-Band Compression," JASA 28, 573, 1956.

(4) S. S. L. Chang, "On the Filter Problem of the Power-Spectrum Analyzer," IRE 42, 1278, 1954.

(5) Cyril M. Harris and William M. Waite, "Response of Spectrum Analyzers of the Bank-of-Filters Type to Signals Generated by Vowel Sounds," JASA 35, 1972, 1963.

(6) D. Gabor, "Theory of Communication," JIEE Part III 93, 429, 1946.

(7) J. Klapper and C. M. Harris, "On the Response and Approximation of Gaussian Filters," IRE Trans. on AUDIO AU-7, 80, 1959.

(8) Cyril M. Harris and William M. Waite, "Gaussian-Filter Spectrum Analyzer," JASA 35, 447, 1963.

(9) R. M. Fano, "Short-Time Autocorrelation Function and Power Spectra," JASA 22, 546, 1950.

(10) M. R. Schroeder and B. S. Atal, "Generalized Short-Time Power Spectra and Autocorrelation Functions," JASA 34, 1679, 1962.

(11) J. Capon, "High-Speed Fourier Analysis with Recirculating Delay-Line-Heterodyner Feedblack Loops," IRE Trans. on Instrumentation I-10, 32, 1961.

(12) A. Michael Noll, 'Short-Time Spectrum and "Cepstrum" Techniques for Vocal-Pitch Detection,' JASA 36, 296, 1964.

(13) J. Liljencrants, "51-Channel Analyzer for Spectrum Sampling" STL-QPSR*-1/1962, p3.

(14) R. K. Potter, G. A. Kopp and H. C. Green, "Visible Speech", D. van Nostrand, New York, 1947.

(15) G. FANT, "Acoustic Analysis and Synthesis of Speech with Applications to Swedish," Ericsson Technics No. 1, 1959.

(16) J. L. Flanagan, "Models for Approximating Basilar Membrane Displacement," BSTJ 39, 1163, 1960.
(and the Part II, BSTJ 41, 959, 1962.)

(17) J. Liljencrants, "RASSLAN-a 6-channel closed loop sectioning device," STL-QPSR*-2/1960, p1.

(18) T. Sakai and S. Doshita, "Conversion of sound spectrum for digital display," Record of the 1960 Joint Convention of the IECEJ** and others.

(19) T. Sakai and S. Doshita, "Analysis of speech sound by means of single tuned filter," Record of the 1965 Joint Convention of the IECEJ** and others.

(20) George A. Hellwarth, "An Automatic Speech Formant Tracking Filter," The University of Michigan, Communication Science Laboratory Report No. 10.

(21) K. Maeda, T. Sakai and S. Doshita, "Analysis of Semi-vowel in Monosyllable," Record of the 1959 Convention of the IECEJ**

(22) K. Maeda, T. Sakai and S. Doshita, "Analysis of Speech Sound by means of Sequential Circuit," Record of the 1959 Joint Convention of the IECEJ** and others.

(23) John M. Heintz and Kennes N. Stevens, "On the Properties of Voiceless Fricative Consonants," JASA 33, 589, 1961.

(24) George W. Hughes and Morris Halle, "Spectral Properties of Fricative Consonants," JASA 28, 303, 1956.

(25) J. M. Heinz, "Analysis of Fricative Consonants," Quarterly Progress Report No. 60, Research Laboratory of Electronics M. I. T. Jan. 15, 1961, pp 181--184.

(26) Toshio Sato, "On the Differences in Time Structures of Voiced and Unvoiced Stop consonants," Journal of Acous. Soc. of Japan 14, 117, 1958.

(27) R. Jakobson, C. G. Fant and M. Halle, "Preliminaries to Speech Analysis," Tech. Rep. No. 13, Acous. Lab. MIT., Jan. 1952.

(28) M. Halle, G. M. Hughes and J. P. A. Radley, "Acoustic Properties of Stop Consonants," JASA 29, 107, 1957.

(29) O. Fujimura, "Spectra of Nasalized Vowels," Quarterly Progress Report No. 58 Research Laboratory of Electronics M. I. T. 1960, pp 214--218.

(30) O. Fujimura, "Analysis of nasal consonants," Quarterly Progress Report No. 60, Research Laboratory of Electronics M. I. T. Jan. 15, 1961, pp 184--188.

(31) For example, Shiro Hattori,"音声学(phonetics)" p 55, Iwanami Shoten, 1951.

(32) K. Nakata, "Synthesis and Perception of Nasal Consonants," JASA 31, 6, 661, 1959.

(33) Shiro Hattori, Kengo Yamamoto and Osamu Fujimura, "Nasal Consonants and Nasalized Vowels," Jour. of the Acous. Soc. of Japan 12, 197, 1956.

(34) H. Takahashi et al., "Studies on the Movement of the Nasopharyngeal Wall Related to Speech," STUDIA PHONOLOGICA II, 47, 1962.

(35) C. E. Shannon, "A Mathematical Theory of Communication," BSTJ 27, 379 and 632, 1948.

(36) C. E. Shannon, "Prediction and Entropy of Printed English," BSTJ 30, 50, 1951.

(37) P. B. Denes, "On the statistics of Spoken English," JASA 35, 892, 1963.

(38) Paul G. Hoel, "Introduction to Mathematical Statistics," John Wiley & Sons, 1947.

(39) J. E. Shoup, "Phoneme Selection for Studies in Automatic Speech Recognition," JASA 34, 397, 1962.

(40) T. Sakai and S. Doshita "Statistics of the Japanese Phoneme Sequences," Record of the 1963 Joint Convention of the IECEJ** and others.

(41) T. Sakai and S. Doshita, "Speech Recognition System of Conversational Sounds," Jour. of IECEJ,** 46, 1696, 1963.

(42) Caldwell P. Smith, "A Phoneme Detector," JASA 23, 446, 1951.

*STL-QPSR: Speech Transmission Labolatory Quarterly Progress and Status Report, Royal Institute of Technology.

**IECEJ: The Institute of Electrical Communication Engineers of Japan.

REFERENCES OF PART II

(1) K. H. Davis, R. Biddulph and S. Balashek,"Automatic Recognition of Spoken Digits," JASA 24, 637, 1952.

(2) E. E. David, Jr., "Artificial Auditory Recognition in Telephony," IBM Journal 2, 294, 1958.

(3) H. Dudley and S. Balashek,"Automatic Recognition of Phonetic Patterns in Speech," JASA 30, 721, 1958.

(4) P. Denes and M. V. Mathews, "Spoken Digit Recognition Using Time-Frequency Pattern Matching," JASA 32, 1450, 1960.

(5) Y. Kato, S. Chiba and K. Nagata, "Spoken Digit Recognizer," Jour. IECEJ* 47, 1319, 1964.

(6) H. F. Olson and H. Belar, "Phonetic Typewriter," JASA 28, 1072, 1956.

(7) D. B. Fry "Theoretical aspects of mechanical speech recognition," Jour. Brit. IRE 19, 211, 1959.

(8) P. Denes,"The Design and Operation of Mechanical Speech Recognizer at University College London," Jour. Brit. IRE 19, 219, 1959.

(9) T. Sakai and S. Doshita, "Speech Recognition System of Conversational Sounds," Jour. IECEJ* 46, 1696, 1963.

(10) J. C. R. Licklider and I. Pollack, "Effects of Differentiation, Integration and Infinite Peak Clipping upon the Intelligibility of Speech," JASA 20, 42, 1948.

(11) J. C. R. Licklider, "The Intelligibility of Amplitude-Dichotomized Time-Quantized Speech Waves," JASA 22, 820, 1950.

(12) E. Peterson, "Frequency Detection and Speech Formants," JASA 23, 668, 1951.

(13) P. Marcou and J. Daguet,"New Methods of Speech Transmission" Information Theory, Third London Symposium, p231, 1955.

(14) E. Colin Cherry and V. J. Phillips, "Some Possible Uses of Single Sideband Signals in Formant-Tracking Systems," JASA 33, 1067, 1961.

(15) Keiji Hiramatsu,"Zero-crossing Information of S. S. B. Speech Signal" Jour. of Acous. Soc. of Japan, 18, 301, 1962.

(16) Keiji Hiramatsu and Yukinobu Kumakawa, "Speech Band Compression System Using S. S. B. - Clipping——Formac——," Jour. of Acous. Soc. of Japan, 18, 310, 1962.

(17) J. L. Daguet, "Speech Compression CODIMEX System" IEEE Trans. on AUDIO,

AU-11, 63, 1963.

(18) C. R. Howard, "Speech Analysis-Synthesis Scheme Using Continuous Parameters," JASA, 28, 1091, 1956.

(19) George A. Hellwarth, "An Automatic Speech Formant Tracking Filter" Report No. 10, Communication Sciences Laboratory, The University of Michigan, 1962.

(20) O. Rice, "Mathematical Analysis of Random Noise," BSTJ 24, 46, 1945.

(21) A. J. Rainal, "Zero-Crossing Intervals of Gaussian Process," IRE Trans. on Information Theory, IT-8, 372, 1962.

(22) W. B. Davenport Jr., "A Study of Speech Probability Distributions," MIT Research Laboratory of Electronics, Technical Report No. 148, August 25, 1950.

(23) T. Sakai and S. Inoue, "New Instruments and Methods for Speech Analysis," JASA, 32, 441, 1960.

(24) T. Sakai and S. Doshita, "The Automatic Speech Recognition System for Conversational Sound", IEEE Trans. on Electronic Computers, EC-12, 835, 1963.

(25) T. Sakai, S. Doshita and K. Hashimoto, "The Automatic Recognition System of the Japanese Monosyllable," Technical Report of the Professional Group on Automaton and Autonatic Control of the IECEJ*, January, 1961.

(26) R. K. Potter & J. C. Steinberg, "Toward the Specification of Speech" JASA, 22, 807, 1950.

(27) Kazuo Kondo, "Preliminary to the Mathematical phonetics" University of Tokyo Press, 1964.

(28) K. N. Stevens, "Toward a Model of Speech Recognition", JASA, 31, 1490, 1959.

(29) S. Inoue, "Study on the Properties of Japanese Speech Sound and its Discrimination" Doctor thesis submitted to Kyoto University, 1959.

(30) T. Sakai and S. Doshita, "An Automatic Speech Recognition System", Transaction on Automaton and Automatic Control of the IECEJ; No.2, 1962.

(31) T. Sakai, S. Doshita, K. Nagata and T. Sekimoto, "Phonetic Typewriter", Proceedings of Speech Communication Seminar, Stockholm, 1962.

(32) T. Sakai and S. Doshita, "The Phonetic Typewriter", Proceedings of IFIP Congress 62, Munich.

(33) T. Sakai and S. Doshita, "Segmentation of Connected Speech", Record of the 1963 Convention of the Information Processing Society of Japan.

IECEJ*: The Institute of Electrical Communication Engineers of Japan.

REFERENCES OF PART II

(1) K. H. Davis, R. Biddulph and S. Balashek,"Automatic Recognition of Spoken Digits," JASA 24, 637, 1952.

(2) E. E. David, Jr., "Artificial Auditory Recognition in Telephony," IBM Journal 2, 294, 1958.

(3) H. Dudley and S. Balashek,"Automatic Recognition of Phonetic Patterns in Speech," JASA 30, 721, 1958.

(4) P. Denes and M. V. Mathews, "Spoken Digit Recognition Using Time-Frequency Pattern Matching," JASA 32, 1450, 1960.

(5) Y. Kato, S. Chiba and K. Nagata, "Spoken Digit Recognizer," Jour. IECEJ* 47, 1319, 1964.

(6) H. F. Olson and H. Belar, "Phonetic Typewriter," JASA 28, 1072, 1956.

(7) D. B. Fry "Theoretical aspects of mechanical speech recognition," Jour. Brit. IRE 19, 211, 1959.

(8) P. Denes, "The Design and Operation of Mechanical Speech Recognizer at University College London," Jour. Brit. IRE 19, 219, 1959.

(9) T. Sakai and S. Doshita, "Speech Recognition System of Conversational Sounds," Jour. IECEJ* 46, 1696, 1963.

(10) J. C. R. Licklider and I. Pollack, "Effects of Differentiation, Integration and Infinite Peak Clipping upon the Intelligibility of Speech," JASA 20, 42, 1948.

(11) J. C. R. Licklider, "The Intelligibility of Amplitude-Dichotomized Time-Quantized Speech Waves," JASA 22, 820, 1950.

(12) E. Peterson, "Frequency Detection and Speech Formants," JASA 23, 668, 1951.

(13) P. Marcou and J. Daguet,"New Methods of Speech Transmission" Information Theory, Third London Symposium, p231, 1955.

(14) E. Colin Cherry and V. J. Phillips, "Some Possible Uses of Single Sideband Signals in Formant-Tracking Systems," JASA 33, 1067, 1961.

(15) Keiji Hiramatsu,"Zero-crossing Information of S. S. B. Speech Signal" Jour. of Acous. Soc. of Japan, 18, 301, 1962.

(16) Keiji Hiramatsu and Yukinobu Kumakawa, "Speech Band Compression System Using S. S. B. - Clipping——Formac——," Jour. of Acous. Soc. of Japan, 18, 310, 1962.

(17) J. L. Daguet, "Speech Compression CODIMEX System" IEEE Trans. on AUDIO,

AU-11, 63, 1963.

(18) C. R. Howard, "Speech Analysis-Synthesis Scheme Using Continuous Parameters", JASA, 28, 1091, 1956.

(19) George A. Hellwarth, "An Automatic Speech Formant Tracking Filter" Report No. 10, Communication Sciences Laboratory, The University of Michigan, 1962.

(20) O. Rice, "Mathematical Analysis of Random Noise", BSTJ 24, 46, 1945.

(21) A. J. Rainal, "Zero-Crossing Intervals of Gaussian Process", IRE Trans. on Information Theory, IT-8, 372, 1962.

(22) W. B. Davenport Jr., "A Study of Speech Probability Distributions", MIT Research Laboratory of Electronics, Technical Report No. 148, August 25, 1950.

(23) T. Sakai and S. Inoue, "New Instruments and Methods for Speech Analysis", JASA, 32, 441, 1960.

(24) T. Sakai and S. Doshita, "The Automatic Speech Recognition System for Conversational Sound", IEEE Trans. on Electronic Computers, EC-12, 835, 1963.

(25) T. Sakai, S. Doshita and K. Hashimoto, "The Automatic Recognition System of the Japanese Monosyllable", Technical Report of the Professional Group on Automaton and Autonatic Control of the IECEJ*, January, 1961.

(26) R. K. Potter & J. C. Steinberg, "Toward the Specification of Speech" JASA, 22, 807, 1950.

(27) Kazuo Kondo, "Preliminary to the Mathematical phonetics" University of Tokyo Press, 1964.

(28) K. N. Stevens, "Toward a Model of Speech Recognition", JASA, 31, 1490, 1959.

(29) S. Inoue, "Study on the Properties of Japanese Speech Sound and its Discrimination" Doctor thesis submitted to Kyoto University, 1959.

(30) T. Sakai and S. Doshita, "An Automatic Speech Recognition System", Transaction on Automaton and Automatic Control of the IECEJ*, No.2, 1962.

(31) T. Sakai, S. Doshita, K. Nagata and T. Sekimoto, "Phonetic Typewriter", Proceedings of Speech Communication Seminar, Stockholm, 1962.

(32) T. Sakai and S. Doshita, "The Phonetic Typewriter", Proceedings of IFIP Congress 62, Munich.

(33) T. Sakai and S. Doshita, "Segmentation of Connected Speech", Record of the 1963 Convention of the Information Processing Society of Japan.

IECEJ*: The Institute of Electrical Communication Engineers of Japan.